Fall 2020

Fuzzy logistic regression for detecting differential DNA methylation regions

Tarek M. Bubaker Bennaser

Department: Mathematics and Statistics

Recommended Citation

Bennaser, Tarek M. Bubaker, "Fuzzy logistic regression for detecting differential DNA methylation regions" (2020). *Doctoral Dissertations*. 2938.
https://scholarsmine.mst.edu/doctoral_dissertations/2938

FUZZY LOGISTIC REGRESSION FOR DETECTING DIFFERENTIAL DNA METHYLATION REGIONS

by

TAREK M. BUBAKER BENNASER

A DISSERTATION

Presented to the Graduate Faculty of the

MISSOURI UNIVERSITY OF SCIENCE AND TECHNOLOGY

In Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

in

MATHEMATICS WITH STATISTICS EMPHASIS

2020

Approved by

Dr. Gayla R . Olbricht, Advisor
Dr. V. A. Samaranayake
Dr. Xuerong Wen
Dr. Yanzhi Zhang
Dr. Ronald L. Frank

ABSTRACT

Epigenetics is the study of changes in gene activity or function that are not related to a change in the DNA sequence. DNA methylation is one of the main types of epigenetic modifications, that occur when a methyl chemical group attaches to a cytosine on the DNA sequence. Although the sequence does not change, the addition of a methyl group can change the way genes are expressed and produce different phenotypes. DNA methylation is involved in many biological processes and has important implications in the fields of biomedicine and agriculture.

Statistical methods have been developed to compare DNA methylation at cytosine nucleotides between populations of interest (e.g., healthy and diseased) across the entire genome from next generation sequence (NGS) data. Testing for the differences between populations in DNA methylation at specific sites is often followed by an assessment of regional difference using post hoc aggregation procedures to group neighboring sites that are differentially methylated. Although site-level analysis can yield some useful information, there are advantages to testing for differential methylation across entire genomic regions. Examining genomic regions produces less noise, reduces the numbers of statistical tests, and has the potential to provide more informative results to biologists.

In this research, several different types of logistic regression models are investigated to test for differentially methylated regions (DMRs). The focus of this work is on developing a fuzzy logistic regression model for DMR detection. Two other logistic regression methods (weighted average logistic regression and ordinal logistic regression) are also introduced as alternative approaches. The performance of these novel approaches are then compared with an existing logistic regression method (MAGIg) for region-level testing, using data simulated based on two (one plant, one human) real NGS methylation data sets.

ACKNOWLEDGMENTS

I would like to give special thanks to my advisor, Dr. Gayla R. Olbricht, for her effort, guidance, patience, kindness, and valuable suggestions during my work with her. Also, I would like to thank Dr. V.A. Samaranayake for his help and support during my Ph.D. In addition, I thank my committee members, Dr. Yanzhi Zhang, Dr. Xuerong Wen, and Dr. Ronald L. Frank. Special thanks, to my children, Ahmed, Rinad, Layaan, Zakiya, and Allen, and to my wife, without her patience it would have been impossible for me to complete my Ph.D. studies.

TABLE OF CONTENTS

Page

LIST OF ILLUSTRATIONS

Figure Page

LIST OF TABLES

Table Page

1. INTRODUCTION

1.1. BASIC CONCEPTS OF GENETICS AND EPIGENETICS

This dissertation focuses on statistical methodology for analyzing epigenetic data. Specifically, statistical methods that aim to identify differentially methylated genomic regions are addressed in this work. In this section, basic concepts of genetics and epigenetics are provided as background to understand the biological nature of the topic.

1.1.1. Genetics. Genetics is the scientific study of inherited variation. Genes are sections of deoxyribonucleic acid (DNA) located inside each cell of an organism that are the fundamental units of heredity and play a role in determining the nature of living organisms (Lim and Maher, 2010). Organisms inherit phenotypes (characteristics) based on genotypes (the version of genes an individual possesses). For example, hair color, eye color, height, and the sound of a person's voice are phenotypes, the outward manifestations of the genotypes. It is of interest to determine the connection between genotype and phenotype. One way to investigate this connection at the molecular level involves the field of molecular genetics, a field that combines genetics with molecular biology. The Central Dogma of Molecular Biology (Varley *et al.*, 2013) offers a way to understand how genes are transferred to functional information (Lim and Maher, 2010; Varley *et al.*, 2013). The Central Dogma of Molecular Biology describes this two-step process (transcription and translation) that takes the information in DNA and uses it to produce proteins: DNA→ RNA→ protein (DeBruyn, 2012). That is, the information contained in genes (DNA) is transcribed to messenger RNA, which is then translated to proteins. Proteins are the essential functional units of the cell.

The field of genomics involves studies that are conducted at the genome level, the study of the complete set of DNA in an organism (Varley *et al.*, 2013). DNA is double-stranded and is comprised of millions of nucleotides. Nucleotides function as subunits and are composed of a nitrogen nucleobase (i.e., cytosine(C), guanine(G), adenine(A), and thymine(T)), a five-carbon sugar, and at least one phosphate group (Lodish *et al.*, 2000). These bases can be categorized into two classes: the purine bases (A and G) that have a double ring structure and the pyrimidine bases (T and C) that only have a single ring. On the double stranded DNA (dsDNA), a purine on one strand pairs with a pyrimidine on the other strand. This is called complementary base pairing and it always occurs where A is paired with T while G is paired with C. The length of the dsDNA is measured by the number of base pairs (bp) of the product. The sequence length ranges from a few thousand (i.e., kilo base pairs (kbp)) in single-celled organisms to several million (i.e., mega base pairs (Mbp)) per molecule for complex organisms (Topal and Fresco, 1976).

DNA molecules are tightly packed around histone proteins to make structures called chromosomes. In humans there are 23 pairs of chromosomes. The largest chromosome, chromosome 1, contains about 2000 genes. The smallest chromosome, chromosome 21, contains about 200-300 genes. The mapping of genomes for particular organisms enables scientists to identify and record the location of genes within the DNA, the distances between genes on a chromosome, and gene functions. Mapping genomes allows the DNA to be annotated into important structures such as protein-coding genes (a sequence of DNA that codes for protein) and non-coding sequences (which do not provide instructions for making proteins). The map of these structures can be further annotated into additional units such as exons, introns, CpG islands and promoter regions that often play specific roles in different molecular processes.

The Human Genome Project (Venter *et al.*, 2001) was an international genome-wide study of the DNA sequence in humans with the goals of identifying the sequence of the approximately three billion nucleotide positions that make up human DNA and identifying the locations of the genes within this sequence. The order of nucleotides in a gene is revealed by DNA sequencing. Initially, sequence information was used in organism specific genome projects such as the Human Genome Project to map the position of a gene or genetic marker to be used as a reference for future studies of that organism. With advances in technology, whole genome sequencing can now be used to reveal all of the genetic information from any individual. Such genome-wide studies are powerful tools for exploring genetic contributions to phenotypic variation and have the potential to allow for important health advances, such as personalized medicine, in the future.

1.1.2. Epigenetics. Epigenetics is the study of biological mechanisms that will switch genes on and off without any change in the DNA sequence. The word "epigenetics" comes from the Greek root “epi", which means “above" or “on top of" the genetic information. Epigenetics encompasses all the information contained within the cell and expressed for more than one cell generation, as the DNA sequence remains stable (Vaillant and Paszkowski, 2007). In other words, epigenetics is the study of heritable changes in gene expression (active versus inactive genes) that cannot be explained by changes in the DNA sequence (Russo *et al.*, 1996). DNA methylation and histone modifications are two key epigenetic mechanisms (Figure 1.1) (Jones and Baylin, 2007). Both of these mechanisms involve the addition of chemical marks to the DNA or histone proteins. DNA methylation occurs when a methyl (Me) group attaches to a cytosine (C) base on the DNA molecule and is the focus of this work. Histone modifications occur when certain chemical groups (e.g., methyl, acetyl) attach to the tails of histone proteins. Epigenetic modifications help to control gene expression (Jaenisch and Bird, 2003; Vaillant and Paszkowski, 2007; Zilberman *et al.*, 2007), and epigenetic aberrations correlate with cancer (Feinberg and Tycko, 2004; Feinberg and Vogelstein, 1983) and other diseases (Robertson, 2005; Shames *et al.*, 2007). Environmental and lifestyle factors may influence epigenetic mechanisms, such as DNA methylation and histone modifications. An advantage of these environmental influences is that drugs can be developed to modify epigenetic patterns in cancer cells (Issa and Kantarjian, 2009).

1.1.3. DNA Methylation. DNA methylation is an epigenetic mechanism that occurs by adding a methyl (Me) group to DNA, thereby often modifying the function of the genes and affecting gene expression (Bird, 2002; Suzuki and Bird, 2008). In mammals,

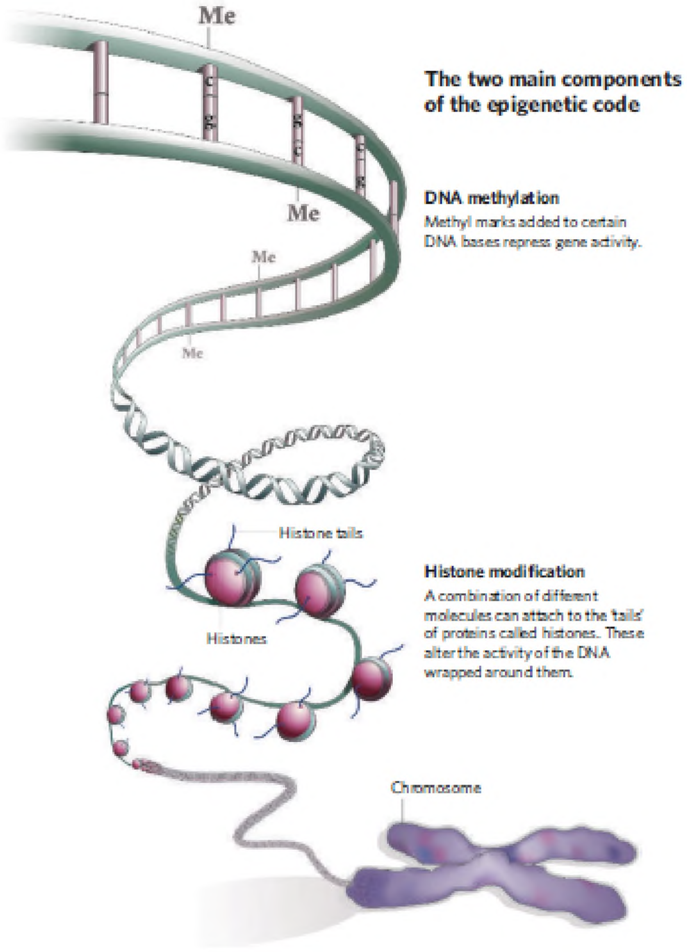
Me
The two main components of the epigenetic code
DNA methylation
Methyl marks added to certain DNA bases repress gene activity.
Me
Me
Me
Histone tails
Histone modification
A combination of different molecules can attach to the 'tails' of proteins called histones. These alter the activity of the DNA wrapped around them.
Histones
Chromosome

Figure 1.1. Two key epigenetic modifications: (1) DNA methylation and (2) histone modification. Image from Qiu (2006).

DNA methylation occurs primarily when cytosine (C) is followed by guanine (G) in the 5’-3’ direction of the DNA sequence; although there is increasing evidence that methylation also occurs in other contexts. This is denoted as CG or CpG, the latter notation showing that cytosine and guanine are connected by a phosphate on one of the DNA strands (Li and Zhang, 2014). In plants, DNA methylation occurs in the contexts of: CG, CHG and CHH (where H = A, T or C). At least three DNA methylation pathways exist in plants, and each pathway appears to methylate cytosines in different sequence contexts (Law and Jacobsen, 2010). In mammals, DNA methylation patterns are established and maintained by three

DNA methyltransferases: DNMT3a, DNMT3b and DNMT1 (Margot *et al.*, 2003). Plants also have methyltransferase enzymes, some similar to DNMT1 and others unique to plants (Law and Jacobsen, 2010). CpG islands are short stretches of DNA in which the frequency of the CG sequence is higher than other regions. Usually, CpG islands occur upstream of many genes that are unmethylated (Law and Jacobsen, 2010).

Recent research suggests that the relationship between genetic variation, DNA methylation and expression is complex (van Eijk *et al.*, 2012). DNA methylation is important for many biological processes, including genomic imprinting, X-chromosome inactivation, embryonic development and the silencing of transposable elements (Bird, 2002; Finnegan, 2010; Gehring and Henikoff, 2007; Kim *et al.*, 2009; Slotkin and Martienssen, 2007). In plants, DNA methylation is important for genome stability and plant improvement (Finnegan, 2010; Gehring and Henikoff, 2007). In humans, particular DNA methylation patterns have been associated with the development of cancer (Feinberg and Vogelstein, 1983). A general loss of DNA methylation (hypomethylation) that occurs with an increase in methylation (hypermethylation) at the CpG islands in promoter regions is often found in cancer cells (Feinberg and Vogelstein, 1983; Shames *et al.*, 2007).

1.2. NEXT-GENERATION SEQUENCING TECHNOLOGY

Next-generation sequencing (NGS) technologies are rapidly becoming an integral part of genetic research and discovery. Next-generation sequencing is also known as high-throughput sequencing and is a catch-all term used to describe a number of different modern sequencing technologies. NGS techniques combine high-throughput capability and high sequencing read accuracy with a low cost per base. The cost for the sequencing of an entire human genome has dropped from about $10 million to about $1000 in just the last 15 years (Brettin *et al.*, 2015). NGS methods offer advantages for such large-scale studies over the traditional Sanger sequencing developed in 1977 as they allow DNA and RNA to be sequenced much more quickly and cheaply.

NGS technology has enabled researchers to conduct genome-wide investigations of many different molecular level phenomena, including gene expression and epigenetic modifications such as DNA methylation. Next-generation sequencing platforms can deliver a great amount of useful DNA methylation information at individual cytosines due to their higher accuracy and sensitivity, making them suitable for epigenomic investigations (Berglund *et al.*, 2011). Several different companies manufacture NGS technologies (e.g., Illumina, Roche 454, Life Technologies) and their specific technical details may differ. However, there is a set of general processing steps shared between them that is illustrated in Figure 1.2 (Voelkerding *et al.*, 2010). These common NGS steps include library preparation, amplification, sequencing, imaging and alignment, which result in millions of sequencing reads per run. Coverage or sequencing depth is one of the common measures of the amount of sequence data generated, and it refers to the average number of times each base in the genome is sequenced (Voelkerding *et al.*, 2010)

1.2.1. Genome-wide Methylation Profiling Approaches. DNA methylation can be investigated at a genome-wide level using a variety of technologies, including microarrays and NGS methods. These technologies provide an opportunity to rapidly analyze the genome of any organism. The advantage of NGS technologies is that they can cover cytosine sites across the entire genome and not just a pre-chosen subset of locations covered by microarrays. Next-generation technologies can also cover a wide range of the genome including repetitive elements (Lister *et al.*, 2008).

Novel approaches have been developed with NGS to obtain genome-wide profiles of DNA methylation. Some methods require bisulfite-converted genomic DNA, such as MethylC-seq and RRBS (reduced representation bisulfite sequencing) (Lister *et al.*, 2008; Meissner *et al.*, 2008), some rely on the enrichment of methylated DNA, such as MeDIP-seq (methylated DNA immunoprecipitation sequencing) and MBD-seq (methylated DNA binding domain sequencing) (Down *et al.*, 2008; Serre *et al.*, 2009) and some use methylation-sensitive characteristics of restriction enzymes to digest genomic DNA. Each method has

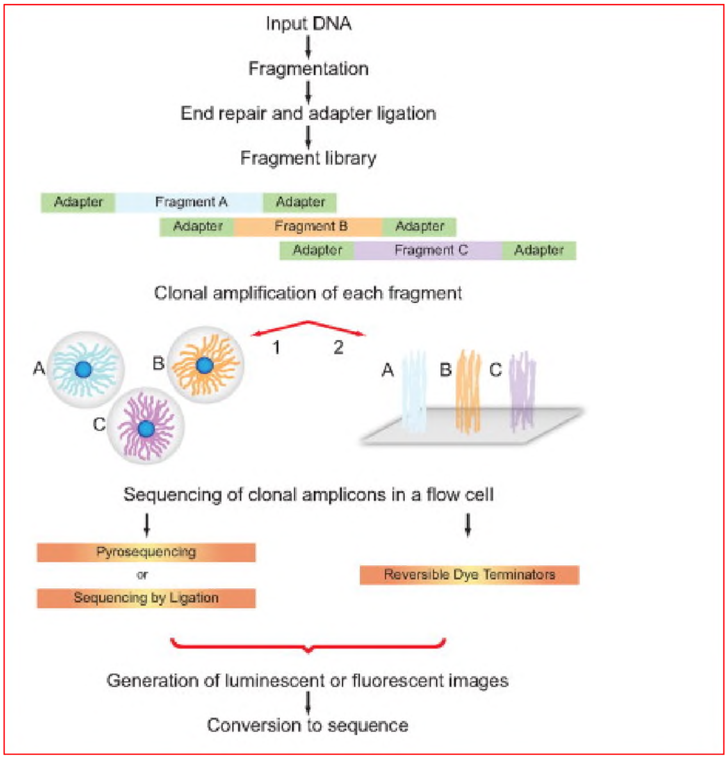
Input DNA
Fragmentation
End repair and adapter ligation
Fragment library
Adapter
Fragment A
Adapter
Adapter
Fragment B
Adapter
Adapter
Fragment C
Adapter
Clonal amplification of each fragment
1
2
A
B
C
A
B
C
Sequencing of clonal amplicons in a flow cell
Pyrosequencing
or
Sequencing by Ligation
Reversible Dye Terminators
Generation of luminescent or fluorescent images
Conversion to sequence

Figure 1.2. Next-generation sequencing process steps for platforms requiring clonally amplified templates (Roche 454, Illumina and Life Technologies). Input DNA is converted to a sequencing library by fragmentation, end repair and ligation to platform specific oligonucleotide adapters. Individual library fragments are clonally amplified by either (1) water in oil bead-based emulsion PCR (Roche 454 and Life Technologies) or (2) solid surface bridge amplification (Illumina). Flow cell sequencing of clonal templates generates luminescent or fluorescent images that are algorithmically processed into sequence reads. Image from Voelkerding *et al.* (2010).

advantages and disadvantages with regard to sequencing depth, covered regions, accuracy and cost. For example, MeDIP-seq and MBD-seq cannot provide base pair-specific profiles, but they can reflect high to medium methylation of DNA sequences covering broader regions (Bird, 2002). In contrast, whole genome bisulfite sequencing (WGBS) methods, such as methylC-seq, provide a direct measure of CpG methylation at predefined regions and are considered the gold standard, but this technique is very costly (Suzuki and Bird, 2008). As a compromise, RRBS combines the use of restriction enzymes with NGS and bisulfite sequencing to obtain methylation levels at individual cytosines in a subset of the genome with high CpG content. This reduces the cost at the expense of missing some information in some regions. This works focuses on developing statistical methods for bisulfite-based NGS methods. The bisulfite sequencing technique is reviewed in the next section followed by a detailed description of the RRBS workflow, due to its aforementioned advantages.

1.2.2. Bisulfite Sequencing Methods to Profile DNA Methylation. Bisulfite sequencing is a technique that enables the measurement of DNA methylation percentage at the individual-base level (Frommer *et al.*, 1992). This technique utilizes a process called bisulfite transformation of genomic DNA, in which the DNA molecules undergo a bisulfite treatment that allows methylated cytosines to be differentiated from unmethylated cytosines at the individual-base resolution. By this method, unmethylated CpG sites can lose an amine group to yield the uracil base, whereas methylated cytosines remain unmodified (Frommer *et al.*, 1992). A polymerase chain reaction (PCR) is used to amplify bisulfite-treated DNA. After this is completed, the unmethylated cytosines appear as thymines (Figure 1.3). Therefore, when combined with NGS, full genome methylation profiles can be produced at the individual-base resolution. CpG sites that are methylated will read as a cytosine; whereas unmethylated CpG sites will be read as thymine. The number of methylated reads and unmethylated reads at the individual-base resolution can be counted and converted to methylation percentages at that position.

1.2.3. Reduced Representation Bisulfite Sequencing. Bisulfite sequencing methods are combined with NGS technologies for whole genome studies using methods such as BS-seq or methylC-seq (Masser *et al.*, 2015). Although these WGBS are considered the gold standard, they are expensive and prohibitive for most large-scale projects, especially for large genomes and large sample sizes. Meissner *et al.* (2008) pioneered the reduced representation bisulfite sequencing (RRBS) approach for analyzing and comparing genomic methylation between groups (e.g., case vs. control) with large sample sizes for use in smaller labs (Gu *et al.*, 2010). The RRBS method utilizes bisulfite sequencing with NGS, but it only sequences a subset of the CpG sites rather than the whole genome.

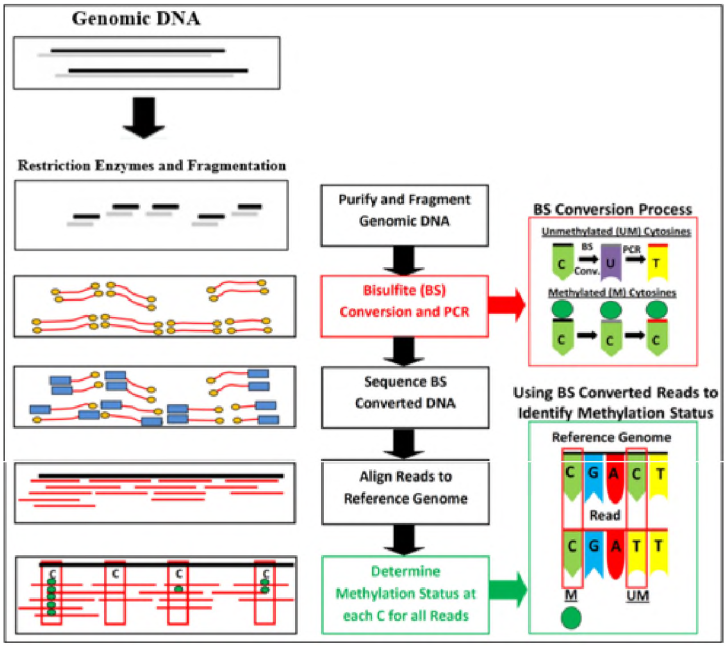
Genomic DNA
Restriction Enzymes and Fragmentation
Purify and Fragment Genomic DNA
Bisulfite (BS) Conversion and PCR
Sequence BS Converted DNA
Align Reads to Reference Genome
Determine Methylation Status at each C for all Reads
BS Conversion Process
Unmethylated (UM) Cytosines
C
BS Conv.
U
PCR
T
Methylated (M) Cytosines
C
C
C
Using BS Converted Reads to Identify Methylation Status
Reference Genome
C G A C T
Read
C G A T T
M
UM
C
C
C
C

Figure 1.3. Workflow of an RRBS library preparation. Image from Olbricht (2012).

RRBS works by using an enzyme that digests genomic DNA for fragments that always start with a C (if the cytosine is methylated) or a T (if the cytosine is not methylated). After this initial step, conversion is completed to establish the methylation levels in DNA for the subset of selected fragments (Raine *et al.*, 2016). Core promoters and CpG islands generated by this fragmentation contain the key regulatory parts of the genome even though RRBS comprises only $\sim$ 1% of the whole genome (Meissner *et al.*, 2005). DNA quantities as little as 10-300 ηg are adequate to produce accurate DNA methylation levels with RRBS (Gu *et al.*, 2011). Therefore, RRBS is suitable for large numbers of clinical samples (e.g., tumors) that only supply a small amount of genomic input DNA material.

The basic steps of preparing an RRBS library (Figure 1.3) are described below. The DNA must first be isolated, so that it can be used to make genomic libraries (Gu *et al.*, 2011). The second step is the digestion reaction and fragmentation. Two commercially available enzymes, MspI and TaqI, are frequently used in this process. These enzymes produce fragments that will contain at least one CpG dinucleotide. MspI and TaqI (Gu *et al.*, 2011) are important to aid in capturing CpG-dense regions while reducing the genomic space. In the third step, the digested and size-selected fragments are bisulfite converted, as described in Section 1.2.2, and then PCR amplified. In step four, all the PCR products that have been amplified from bisulfite-converted fragments are sequenced using an NGS platform. One of the most common RRBS workflows uses the Illumina NGS platform (Krueger *et al.*, 2012). In step five, the short sequence reads are aligned to a reference genome. Finally, in step six, cytosine methylation status is determined across all reads and summarized, resulting in a count of the number of methylated (C) and the number of unmethylated (T) reads at each cytosine sequenced. Note that in mammals, typically only CpG sites are summarized, but all cytosines are of interest in plants.

1.3. LITERATURE REVIEW FOR STATISTICAL METHODS

The ability to measure DNA methylation on a genome-wide scale enables researchers to investigate associations between DNA methylation and different phenotypes and conditions of interest. Identifying locations in the genome that have DNA methylation differences between conditions is referred to as differential methylation testing. These results enable researchers to better understand the role of DNA methylation and locate specific genomic locations that may be promising to target in future investigations. This work specifically focuses on developing improved statistical methods for detecting differentially methylated regions (DMRs) from bisulfite-based NGS data. A review of previous statistical methods that provide tests at the region-level is first given, followed by a brief introduction to one of the primary methods proposed in this work (fuzzy logistic regression). The section ends with a summary of the proposed methods for DMR testing.

1.3.1. Previous Approaches for DMR Testing. Differential methylation testing can be done on an individual site-level or on a region-level. Although there are occasions when researchers are interested studying the relationship between single CpG sites and a phenotype of interest (Weaver *et al.*, 2004), differentially methylated regions are often more predictive features of phenotypes (Lun and Smyth, 2014). Also, differences at any given site may be small and noisy, but variations across a region can often be more easily detected. For region-level testing, it is important to define the region location. Methods that operate on predefined regions must be distinguished from those that define regions of DMRs as part of the method (i.e., the regions are not known in advance). One issue with methods that do not define the region in advance is that the false discovery rate (FDR) needs to be controlled across the regions tested, which is difficult when the number of this is unknown in advance. Therefore, the most straightforward approach is to use predefined regions which can be defined based on annotation (e.g., CpG islands, CpG shores, exons, or introns) or defined by CpG density. One other statistical issue that needs to be considered for DMR detection is that methylation levels between neighboring sites are often highly correlated.

Table 1.1. Statistical methods to detect differentially methylated regions (DMRs).

Method	Design	Region Definition	Statistical Elements	Reference
M3D	RRBS	CpG Density	Kernel-based	Mayo *et al.* (2015)
MAGIg	WGBS or RRBS	Annotation	Fisher's Exact Test, Logistic Regression	Baumann and Doerge (2014)

This work focuses on methods that utilize predefined regions in constructing a region-level test for differential methylation using bisulfite-based NGS data. Although there are many statistical approaches in the literature for differential methylation testing, there are currently only two main methods specifically for detecting DMRs based on predefined regions in bisulfite sequencing data. These methods are summarized in Table 1.1. M3D (Mayo *et al.*, 2015) is a nonparametric statistical method that uses a kernel distance statistic to detect DMRs from predefined regions based on CpG density. Methylation analysis using genome information (MAGI) (Baumann and Doerge, 2014) defines testing regions based on existing annotation information, assumes spatial homogeneity across regions and thus does not adjust for spatial correlations between individual cytosine sites. It proposes two different approaches (MAGIc and MAGIg), of which only MAGIg provides a test at the region-level. MAGIg classifies each cytosine as either methylated or unmethylated based on a prior threshold determined by k-means (k=2) clustering. It performs a single Fisher's exact test (FET) for unreplicated data or a logistic regression for replicated data for each region.

This dissertation builds on the work of MAGIg and focuses on developing methods for DMR testing over predefined regions using different logistic regression methodologies. Specifically, a fuzzy logistic regression analysis technique is developed for DMR detection and two other logistic regression approaches are also proposed. The performance of the different logistic regression approaches are compared to MAGIg, the only other method that utilizes a logistic regression methodology for DMR testing at predefined regions.

1.3.2. A Brief Introduction to Fuzzy Regression. In this work, one of the main new methods proposed is a fuzzy logistic regression (FLR) model. The fuzzy regression modeling approach is introduced here to provide a brief background about this type of modeling. Fuzzy regression models are used to evaluate the functional relationship between a response variable and one or more explanatory variables in a fuzzy environment. A fuzzy environment may occur due to the vague or imprecise nature of the response observations. In the case of a binary response, a logistic regression is often used, but when the response observations are fuzzy the classical logistic regression may not be appropriate. Several methods have been presented to estimate fuzzy regression models such as Tanaka's linear programming approach and the fuzzy least-squares approach by focusing on the extension principle (Tanaka *et al.*, 1982) for minimizing errors between the given outputs and the estimated outputs (Diamond, 1987). Fuzzy regression modeling is used in many areas such as economics, engineering, biology and physical sciences (Chang and Lee, 1996). In this work, a fuzzy logistic regression analysis has been developed to detect DMRs. FLR is an extension of the classical logistic regression analysis in which some elements of the model are represented by fuzzy numbers (Kao and Chyu, 2002). The uncertainty in this type of regression model become fuzziness, not randomness. More details about the FLR proposed for DMR detection are given in Section 2.

1.3.3. Summary of Proposed Methods for DMR Testing. The main purpose of this dissertation is to develop improved methods for detecting differentially methylated regions in bisulfite-based NGS studies with predefined regions. The MAGIg method (Baumann and Doerge, 2014) is one of the primary existing methods used in previous work for this type of data and testing. MAGIg uses a logistic regression approach, but it has some limitations that are explored in this dissertation. Building on the logistic regression framework, this research proposes three new logistic regression approaches for DMR testing and investigates their performance via simulation studies. These three approaches include ordinal logistic regression (OLR), a weighted average logistic regression (WALR) and fuzzy logistic regression (FLR). All three methods are described in further details in Section 2. In particular, it is important to highlight that this work represents a novel application of fuzzy logistic regression procedures for analysis of DNA methylation data from bisulfite-based NGS studies. Specifically, these FLR methods will enable testing DMRs and pinpoint genomic regions that are likely to be biologically meaningful.

A brief summary of the FLR approach is described as follows. The k-means clustering method (k=5) is first used to classify each site into methylation level groups (Very Low, Low, Medium, High and Very High). The mode is then taken across the region to identify a linguistic term. The fuzzy coefficient estimates will be obtained for the model relating condition (e.g., treatment vs. control) to the fuzzy methylation output. A bootstrap-based approach is used to test whether the fuzzy coefficient corresponding to the condition term is zero and a list of significant DMRs is obtained. Finally, the false discovery rate (FDR) is controlled all region-level tests (Benjamini and Hochberg, 1995). Simulation studies will be used to evaluate the performance of FLR compared to MAGIg, OLR and WALR methods. One simulation is based on plant data and another is based on human data, with the goal of investigating model performance in both types of data that are known to differ in DNA methylation structure. A description of the simulation studies and their results are provided in Section 3.

2. METHODS

2.1. INTRODUCTION OF METHODS

The focus of this work is to investigate statistical methods for differential methylation testing across predefined regions using bisulfite sequencing based next-generation sequencing (NGS) data. One of the primary methods for testing over predefined annotation-based regions is the methylation analysis using genome information (MAGIg) approach (Baumann and Doerge, 2014). In this method's genome region-level analysis (MAGIg), each potentially methylated site in a region is first classified as methylated or unmethylated based on a cutoff of the methylation level established through k-means (k=2) clustering. This information is summarized as the fraction of methylated sites across the region, which is then used as the response in a logistic regression model. The predictor variable of interest is often a categorical group variable (e.g., treatment vs. control, disease vs. healthy), although additional predictor variables can be included in the model if they are of interest. Note that for human or mammalian data, CG sites are typically the only sites considered for potential methylation, while in plants CG, CHG, and CHH sites are considered as potentially methylated sites (Law and Jacobsen, 2010). Further references to cytosine sites of interest should be understood to be taken in context of the organism being studied.

The logistic modeling framework is a natural choice for region-level differential methylation testing with NGS data. However, there are several possibilities to consider in improving the logistic regression modeling beyond the MAGIg method. This work explores three alternative methods all within the logistic regression modeling framework to address potential limitations of the MAGIg approach. Specifically, the MAGIg method assumes the methylation status at each site is binary (methylated or unmethylated). However, this may be too restrictive. Although methylation at an individual site in an individual cell may be binary, bulk NGS methods involve analyzing a sample that contains multiple cells, which may not all have the same methylation status at a particular site. This results in the methylation level at each site actually being a proportion of methylated sequencing reads that ranges between 0 and 1 (Baumann and Doerge, 2014).

One alternative approach is to use the weighted methylation level (Schultz *et al.*, 2012) to summarize across the region using the sequencing read information rather than relying on classification into methylated and unmethylated sites. This approach is referred to as weighted average logistic regression (WALR). Another option is to utilize the k-means clustering approach to classify the methylation status of each site, but use k=5 rather than k=2 to correspond to the following methylation status groups: Very Low, Low, Medium, High, Very High. An ordinal logistic regression (OLR) (McCullagh, 1980) can then be implemented for differential methylation testing. Finally, a novel implementation of a fuzzy logistic regression (FLR) (Pourahmad, 2013; Pourahmad *et al.*, 2011) offers a different way of viewing the methylation level of a region as a fuzzy observation.

In this section, the logistic regression framework is first introduced and explained in terms of region-level differential methylation testing via the MAGIg approach. The three alternative methods (WALR, OLR, and FLR) are proposed and presented. A focus is placed on introducing the details of the fuzzy logistic regression modeling approach, since this work is the first time this type of modeling has been proposed for differential methylation testing.

2.2. LOGISTIC REGRESSION

Logistic regression is a type of generalized linear model that is important for modeling categorical response data (Agresti, 1996). It is used in a wide variety of applications, including biomedical studies and genetic data analysis. For example, in biomedical studies, it can be used to model the relationship between the probability of having disease or not and potential risk factors. In genetics, it has been used in quantitative trait loci analysis to model the probability of an offspring inheriting a specific allele as function of different quantitative traits for that offspring (Bird, 2002).

Logistic regression has also been used for testing differential methylation between two groups of interest (e.g., disease vs. healthy) at the cytosine-level (Akalin *et al.*, 2012). This allows researchers to identify potentially important cytosine sites where methylation differences could be playing an important cellular role (e.g., altering gene expression) in disease or other condition being studied. At each cytosine site, a logistic regression model is fit using the observed methylation levels as the response and the condition of interest as the predictor. The methylation level for each cytosine is based on the observed proportion of methylated sequencing reads out of the total number of sequencing reads for the site.

Although site-level information can be informative, researchers are often interested in understanding differences in methylation across genomic regions (Mayo *et al.*, 2015). Region-level differential methylation analysis can be performed in two ways: (1) focus on cytosine-level tests and then summarize across the genomic region or (2) summarize the cytosine-level information first and then test once over the region. This work focuses on the latter since this approach provides a way to test for statistical significance at the region-level. Annotation information or CG density can be used to define the regions of interest. Compared to site-level tests, it also reduces the overall number of tests performed, which is beneficial when addressing multiple testing issues.

In region-level models, a marginalization or summary of the methylation levels across all cytosine sites in the region must be calculated. The MAGIg approach (Baumann and Doerge, 2014)) is one method that implements such a marginalization over the region and then utilizes logistic regression for a region-level test. This dissertation focuses on the benefit of conducting region-level differential methylation tests at predefined regions. Different approaches for marginalizing over the region followed by alternative implementations of logistic regression methodology are investigated. Most notably, the fuzzy logistic regression (FLR) (Pourahmad, 2013; Pourahmad *et al.*, 2011) as an extension to logistic regression is proposed for the first time for identifying differentially methylated regions (DMRs). This section presents the theory of logistic regression and explains how logistic regression differs from conventional regression. A statistical test that is used to assess the significance of individual coefficients for inclusion or exclusion in a logistic regression model is also described.

2.2.1. Logistic Regression and Generalized Linear Models. According to Al-Ghamdi (2002), regression methods are widely used for analysing the relationship between a dependent (response) variable and one or more independent (predictor) variables. One of the most commonly used regression methods is linear regression with a continuous response, which is referred to as conventional regression analysis (CRA). Statistical inferences for this model are applicable if the dependent variable observations are continuous, independent and normally distributed with constant variance and the mean dependent on the independent variable values. When the dependent variable is categorical, CRA is not appropriate since the assumptions needed for inference are not met. The most significant reasons why CRA cannot be used when there is a categorical dependent variable are:

- The dependent variable in CRA should be continuous,
- The dependent variable in CRA should be normally distributed, and
- The dependent variable in CRA could take negative and non-negative values.

These CRA assumptions are not satisfied in cases where the dependent variable is categorical (Al-Ghamdi, 2002). In such cases, logistic regression analysis is often applied (Dayton, 1992). Logistic regression, like CRA, is a statistical technique that is used to explore the relationship between a dependent (response) variable and at least one independent (predictor) variable. The difference is that CRA is used when the dependent variable is continuous, while logistic regression techniques are used with categorical dependent variables. The standard logistic regression model is used when the dependent variable is dichotomous (binary). However, extensions of the logistic regression model have also been developed for polychotomous (multi-category) and ordinal data.

Logistic regression, like other model building techniques, is aimed at finding the best fitting, yet sensible model to assess the relationship between a response variable and at least one independent variable. Logistic regression models are a type of generalized linear model (GLM) that extend on the ideas of CRA models to include response data that are not normally distributed. There are three main components of GLMs: a random component, a systematic component, and a link function. The model component specifies the probability distribution of the response variable (Y). This probability distribution is assumed to be from an exponential family, so it includes the normal distribution but also many other distributions such as binomial and Poisson. This allows the model to accommodate different types of response data such as categorical data that are not normally distributed. The systematic component is the linear combination of the predictor variables ($X's$). The link function is a function of the mean of the response ($E(Y)$) that is set equal to the systematic component in the model (Agresti, 1996).

The general form of a GLM model is written as follows:

$$g(\mu_i) = \sum_{j=0}^{p-1} \beta_j x_{ij}$$

where there are $i = 1, \ldots, n$ observations and $j = 0, \ldots, p-1$ predictors. The μ_i term represents the mean of the response variable, that is $\mu_i = E(Y_i)$, and g is a monotone differentiable function. Here, the random component is the independent $Y's$ that follow some exponential family distribution. The systematic component is $\sum_{j=0}^{p-1} \beta_j x_{ij}$, where typically $x_{ij} = 1$ for all i when $j = 0$ is an intercept term. The link function is $g(\mu_i)$. CRA is actually a special case of the GLM when $g(\mu) = \mu$ and the response follows a normal distribution (Agresti, 1996). The model specification for logistic regression when the response is binary is described in the next section.

2.2.2. Binary Logistic Regression Model. Binary logistic regression is a model that is fitted where there is a dichotomous or binary dependent variable. For DNA methylation research, it is of interest whether a genomic site or region is differentially methylated or not between conditions of interest. Normally the response category of interest is referred to as "success" and typically coded as "1". The other group is known as a "failure" and is coded as "0". In this research, the category of interest will be denoted by a "1" if a site is methylated, otherwise it will be denoted by a "0" if the site is not methylated. The region-level logistic regression model (Agresti, 1996) as formulated by MAGIg for methylation (Baumann and Doerge, 2014) is given as:

$$log\frac{\pi_i}{1-\pi_i} = \alpha + \beta * x_i. \qquad (2.1)$$

A separate model is fit for each region. For methylation data, the response is:

$$y_{is} = \begin{cases} 0 & \text{if the } s^{\text{th}} \text{ site in the genomic region is not methylated for individual } i \\ 1 & \text{if the } s^{\text{th}} \text{ site in the genomic region is methylated for individual } i \end{cases}$$

where $i = 1, 2, \ldots, n$ and $s = 1, 2, \ldots, S$ for a given region. Here, $y_{i\cdot} = \sum_{s=1}^{S} y_{is}$ is the number of methylated sites in the genomic region for individual i, which follows a binomial distribution. The predictor (x_i) is the independent variable, which in this case represents the group of the individual (e.g., treatment vs. control, disease vs. healthy). Note that additional predictors could be added, if of interest. α is the coefficient of the constant term (intercept). β is the slope coefficient of the independent variable. The coefficients (α, β) are estimated via the maximum likelihood (ML) method (Kleinbaum *et al.*, 2008). Note that these coefficients are estimated separately for each region. π_i is the probability of success (methylated) and $1 - \pi_i$ is the probability of failure (not methylated) for the predictor value x_i. The link function in the model is $log(\frac{\pi}{1-\pi})$, which is the log odds or logit function.

According to Kleinbaum *et al.* (2008), logistic regression quantifies the relationship between the dichotomous dependent variable and the predictor(s) using the log odds. The odds indicate the probability that an event ("success") will occur divided by the probability that the event will not happen ("failure") In this work, the odds are $\frac{\pi}{1-\pi}$, which represent the probability that a region will be methylated divided by the probability that a region will not be methylated. Thus, the link function connects the log of the odds to the systematic component of the model in equation (2.1). The odds have a minimum value of zero but no upper limit. A value less than one indicates that an event or success is not likely under those circumstances and a value greater than one indicates that a success is more likely than a failure.

The odds ratio is a ratio of the odds of two different conditions (1 and 2): $\frac{\frac{\pi_1}{1-\pi_1}}{\frac{\pi_2}{1-\pi_2}}$ The odds ratio is also a non-negative value and equals 1 when the odds are the same for the two conditions. When the odds ratio is greater than 1, this indicates that condition 1 is more likely to have a success than condition 2 (vice versa when the odds ratio is less than 1). In the logistic regression model (2.1), it can be seen when both sides of the model equation are exponentiated that the odds are an exponential function of the predictor X. Thus, an interpretation of the coefficient β_g can be made as follows. For every 1 unit increase in X, the odds increase multiplicatively by e^{β}. This means that e^{β} is actually an odds ratio of the odds at $X = x + 1$ divided by the odds at $X = x$ (Agresti, 1996). When X is a binary variable representing a condition of interest (e.g., disease vs. healthy), this simplifies to the odds ratio of the condition represented by a "1" over the condition represented by a "0". The further the odds ratio is from one the stronger the relationship is between X and Y.

2.2.3. Assumptions of Logistic Regression. In CRA, the model is often written as $y = E(Y|X) + \varepsilon$, where ε represents the error term that follows a normal distribution with mean zero and constant variance. Hosmer *et al.* (1997) consider a similar formulation for logistic regression as $y = \pi(x) + \varepsilon$. The logistic regression model assumptions are given below.

- The model error terms ε have a mean of zero and a variance of $\pi(x)(1 - \pi(x))$.
- The conditional distribution of Y given X is binomial with conditional mean $\pi(x)$.
- The conditional mean $\pi(x)$ of the regression equation is between 0 and 1 by using the logit link function.
- Data collected from different individuals are independent.

Hosmer *et al.* (1997) also note that the same principles used when conducting CRA also apply for logistic regression equation with one of the main differences being that the regression will be modeling the log odds of an event or success occurring.

2.2.4. Hypothesis Testing. After estimating the logistic regression model parameters using maximum likelihood estimation, the significance of the independent variable needs to be assessed with regards to predicting the response variable. There are a number of statistical tests that can be used to carry out this assessment and these include the likelihood ratio test, Wald test and Score test (Steyerberg *et al.*, 2001). The Wald test is used in this research. To determine if there is a statistically significant relationship between the predictor and the response the following hypotheses are tested for each region:

$$H_0 : \beta = 0 \ \ vs. \ \ H_1 : \beta \neq 0.$$

The Wald test statistic is:

$$W = \frac{\hat{\beta}^2}{[\widehat{SE}(\widehat{\beta})]^2} \sim \chi^2(1)$$

under H_0 where $\widehat{SE}(\widehat{\beta})$ is the estimated standard error of the parameter estimate $\hat{\beta}$. If the null is rejected at significance level α, then there is a significant relationship between the response and predictor.

Note that testing whether $\beta = 0$ or not corresponds to testing whether the odds ratio (e^{β}) is equal to 1 and represents a test of independence between the response and the predictor variable. For methylation data, this test determines whether there is a statistically significant relationship between the methylation level of a region and the condition of interest. If the null hypothesis is rejected, then the region is said to be differentially methylated.

2.2.5. Logistic Regression with MAGIg. As described in the previous sections, MAGIg (Baumann and Doerge, 2014) employs the logistic regression model (2.1) to test for differentially methylated regions (DMRs) using bisulfite-based NGS data. The regions tested are predefined based on genomic annotation. MAGIg first obtains the methylation level of each cytosine site by calculating the proportion of methylated reads out of the total number of reads for the site. The methylation level of each site is then classified into a binary representation of methylation status (1=methylated, 0=unmethylated) based on a threshold established from k-means clustering (k=2). The clustering is implemented on the observed methylation levels for each chromosome and strand. The threshold is the mean of the two cluster centroids. A marginalization over all sites in the genomic region is then completed to obtain a region-level summary within each group of interest and replicate. This is done by finding the number of methylated and unmethylated sites across all sites in the region. The logistic regression model (2.1) is then fit for each genomic region using the number of methylated cytosines out of the total sites in the region as the response along with the group information (x) as the predictor variable. The Wald test as described in Section 2.2.4 is then used to determine significance of the slope coefficient (β). A p-value from this test is obtained for all genomic regions and the false discovery rate (FDR) is controlled across all of the region-level tests (Benjamini and Hochberg, 1995). Regions whose FDR-adjusted p-value is below the significance level (α) are said to be DMRs.

One drawback of MAGIg is that the binary classification (methylated or unmethylated) at each site does not take into account the sequencing depth (total number of reads) at each site. Since the sequencing depth varies from site-to-site, this ignores the fact that a site that is 20% methylated could arise from a site with 1 out of 5 reads methylated or a site with 20 out of 100 reads methylated. Both would get called as unmethylated if the threshold was, say 40%, and they would both be treated equally when summarizing over the region and finding the number of methylated sites. One alternative that addresses this issue is weighted average logistic regression, which is discussed in Section 2.2.6. Also, a binary classification may not be the most appropriate since some sites have methylation levels that are not close to 0 or 1. One option to address this is to utilize k-means clustering where k=5 to allow for Very Low, Low, Medium, High, and Very High methylation groups. Using this approach an ordinal logistic regression could be fit, which is described in Section 2.3. An alternative way to view this issue is to view the methylation level for the region as a fuzzy observation and employ a fuzzy logistic regression. This approach is described in Section 2.4.

2.2.6. Weighted Average Logistic Regression Analysis. As previously mentioned, the DNA methylation level at each cytosine site represents a survey of the methylation states across the sequencing reads at the site. The site-level methylation calculation is shown in Table 2.1. The numerator represents the total number of cytosine (C_s) reads at site s. The denominator is the total number of reads at the site, which is a sum of the cytosine reads that are methylated and the thymine (T_s) reads that are unmethylated at site s. Note that the methylation status in a single cell would be binary (unmethylated or methylated), but in many studies a sample of cells is needed to obtain enough input DNA for the NGS technology. These types of samples are often referred to as bulk samples and are the type of data investigated in this dissertation. Since some cells may be methylated and others unmethylated at a site, the proportion of methylated reads at a site is the quantity of interest (Schultz *et al.*, 2012).

For region-level testing, a summary over the methylation levels for all sites in a region is needed. One approach is to first classify each site as methylated or unmethylated. This can be done using k-means clustering, as in the MAGIg approach (Baumann and

Doerge, 2014), or by conducting a binomial test at each site (Schultz *et al.*, 2012). After this initial classification is done, the MAGIg method then obtains a region-level methylation value by calculating the fraction of methylated cytosines for the region (Table 2.1, Fraction methylation level). The drawback to this approach is that it does not incorporate information about the individual methylation levels at each site, which range between 0 and 1 rather than being a binary value. Using information about the methylation levels may be important since a change in the proportion of methylated cells at different sites could be indicative of a fundamental phenotypic change (Schultz *et al.*, 2012).

An alternative approach for obtaining region-level methylation values is to calculate the mean methylation level (Table 2.1). This method involves calculating the arithmetic mean of the methylation levels across all sites in the region. That is, the methylation levels at each site are first calculated and the mean of these values is computed across the sites in the region. This does provide the advantage of utilizing the site-level methylation values rather than binary values at each site. However, it does not consider the sequencing depth or coverage (i.e., the total number of unique reads that are present or "cover" a specific site). Sequencing depth varies across sites in the region and the higher the depth the more accurate the estimates of the methylation level (Schultz *et al.*, 2012). The contribution of each site is equal using the arithmetic mean and does not account for this accuracy difference that is present at the different sites.

This dissertation implements an approach for calculating region-level methylation values that addresses the issues of treating the site-level methylation as binary and not accounting varying sequencing depths. The weighted methylation level (Table 2.1) is used to weight the contribution of each site by its sequencing depth in the region-level summary (Schultz *et al.*, 2012). The weighted methylation level is calculated by taking the fraction of methylated reads across all sites in the region over the total methylated and unmethylated reads across all sites in the region. As an example, consider a region with two sites. One site has 70 methylated reads out of a total of 100 reads and the other has 1 methylated read out of a total of 4 reads. The mean methylation level would weight these two values the same and the region-level value would be (0.7+0.25)/2 = 0.475. The weighted mean methylation level would give the site with 100 reads more weight than the site with 4 reads and the calculation would be (70+1)/(100+4) = 0.683. Thus, the mean and weighted methylation levels can be quite different. Figure 2.1 provides an additional example of how the fraction, mean, and weighted methylation level calculations differ for a region with 5 sites. The weighted mean is the method recommended by Schultz *et al.* (2012) since it more widely applicable and addresses the previously mentioned drawbacks.

The weighted mean methylation level can be utilized in the logistic regression model proposed in (2.1) instead of the fraction methylation level used in MAGIg for region-level differential methylation testing. This approach is called the weighted average logistic regression (WALR) throughout this dissertation. Similar to MAGIg, the test for the slope coefficient described in Section 2.2.4 can be used to test for DMRs and the p-value from this test adjusted for multiple testing using the FDR method (Benjamini and Hochberg, 1995). Regions whose FDR-adjusted p-value is below the significance threshold (α) are said to be DMRs.

Table 2.1. Different methods for calculating methylation levels at the site and region levels. Fraction, mean and weighted mean methylation levels are all region-level calculations. C = methylated cytosine, T = unmethylated cytosine, s = position of cytosine, S = total number of cytosine positions in the region, M = an indicator variable that is one when k-means or binomial test identifies the position as methylated (Schultz *et al.*, 2012).

Definition	Calculation
Site methylation level	$C_s/(C_s+T_s)$
Fraction methylation level	$1/S \sum_{s=1}^{S} M_s$
Mean methylation level	$1/S \sum_{s=1}^{S} C_s/(C_s+T_s)$
Weighted methylation level	$\sum_{s=1}^{S} C_s / \sum_{s=1}^{S} (C_s+T_s)$

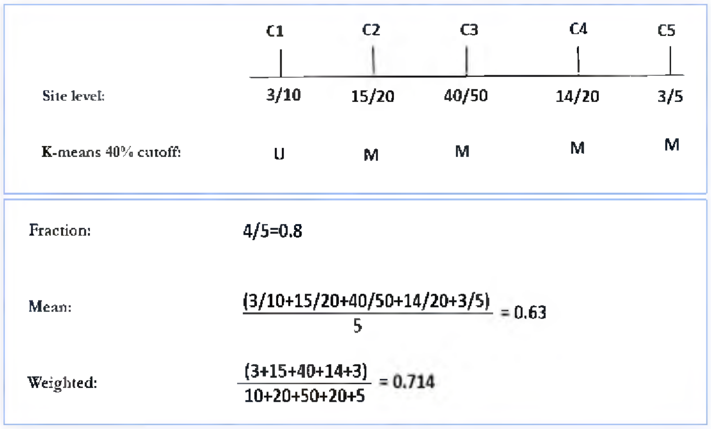

Figure 2.1. Illustration of different methods for calculating methylation levels at sites and regions. An example of the calculations described in Table 2.1 is shown for a region with 5 cytosines. For the k-means 40% cutoff for site-level methylation, the site is considered methylated (M) if > 40% reads are methylated. Otherwise, the site is considered unmethylated (U).

2.3. ORDINAL LOGISTIC REGRESSION

This section introduces an extension of logistic regression analysis to multi-category response data that have an inherent ordering. Many response variables in biostatistics and other fields of study are ordinal in nature, such as cardiac risk levels (low, medium, high) or the grade of a tumor (1, 2, 3). Ordinal logistic regression methods can be used to model the relationship between such an ordinal response variable and a set of explanatory variables. This section describes how ordinal logistic regression can be used for region-level differential methylation testing of NGS data.

Two approaches (MAGIg and WALR) were discussed in Sections 2.2.5 and 2.2.6, respectively, that provided different ways to summarize the methylation level across a region. These region-level methylation values were used as the response in a logistic regression model. Specifically, the MAGIg approach (Baumann and Doerge, 2014) involves classifying each cytosine site as methylated or unmethylated based on a cutoff established from k-means clustering (k=2). However, as previously mentioned, such binary classification has limitations in bulk samples since the data at each site is a mixture of cells, some of which may be methylated and others are not. As an alternative to the approach in MAGIg, this work proposes to classify each site into five ordinal categories for methylation (Very Low, Low, Medium, High, Very High) using k=5 in k-means clustering. A methylation profile over the cytosine sites in a region is calculated by obtaining the number of cytosines in each of the five categories. An ordinal logistic regression is employed to detect differentially methylated regions (DMRs) between two groups or conditions of interest (disease vs. healthy).

This section provides a short review of the ordinal logistic regression (OLR) modeling framework needed for DMR testing. There are different types of logistic regression models for ordinal data, including the cumulative logit, continuation-ratio logit, and the adjacent-category logit models (Agresti, 1996). Each of these models have similar assumptions, but have unique structures for constructing the logit that serves as the link function. The cumulative logit model is chosen in this work for DMR testing. It was proposed by McCullagh (1980) and is a widely used ordinal logistic regression model when there is assumed to be some underlying continuous variable for the response. An overview of the cumulative logit model for ordinal logistic regression is given in the following sections.

2.3.1. Cumulative Logit Model. The proposed ordinal logistic regression model using the cumulative logit model will be employed to identify DMRs within predefined genomic regions. Let the data for the model consist of the observations (x_i, Y_{is}) where x_i is the factor or condition of interest for testing (e.g., treatment or disease status) for $i = 1, \ldots, n$ individuals and Y_{is} is an ordinal observation for the s^{th} site ($s = 1, \ldots, S$) in a given genomic region for individual i. Y_{is} here is an ordinal value (Very Low, Low, Medium, High, Very High) corresponding to the methylation status for a particular site s in the region that is obtained via k-means (k=5) clustering of the methylation levels. When summed over the sites in the region, the counts for each of the categories of Y (subscripts dropped for simplicity) follow a multinomial distribution. Consider modeling the cumulative probability (Agresti, 1996) for Y:

$$logit[p(Y \leq k)] = \pi_1 + \ldots + \pi_k$$

where $k = 1, \ldots, K$ represents the ordinal outcome category and π_k is the probability of methylation status k for a particular region. Here, $K = 5$ and the values of k correspond to the inherent ordering of the methylation categories: 1=Very Low, 2=Low, 3=Medium, 4=High, 5=Very High. This cumulative probability represents the probability that the methylation status for a region falls at or below a particular level k. The logit of this cumulative probability is called the cumulative logit, as defined below (Agresti, 1996):

$$logit[p(Y \leq k)] = log\frac{p(Y \leq k)}{1 - p(Y \leq k)} = log\frac{\pi_1 + \ldots + \pi_k}{\pi_{k+1} + \ldots + \pi_K}.$$

There are a total of $K-1$ of these logits (here, there are 4 cumulative logits). The cumulative logit model for a given region is (Agresti, 1996):

$$logit[p(Y \leq k)] = \alpha_k + \beta x_i \tag{2.2}$$

where α_k is a vector of $K - 1$ intercept parameters and β is the slope regression coefficient that describes the effect of x on the log odds of response category k or below. The parameters are estimated through maximum likelihood estimation. A separate model is fit for each region. The log odds ratio of cumulative probabilities for two values of x (x_1 and x_2) is equal to $\beta(x_1 - x_2)$, which is proportional to the distance between the values of x (Agresti, 1996). McCullagh (1980) called this model a proportional odds model, since each cumulative probability has the same proportionality constant β. Thus, in this proportional odds model, the odds ratio of the event $Y \leq k$ is independent of the category indicator. Note that when $x_1 = 1$ and $x_2 = 0$, as in the case for testing for differential methylation, the log odds ratio of the cumulative probabilities is equal to β and the odds of the response being at or below category k when $x_1 = 1$ is e^β times the odds when $x_1 = 0$.

2.3.2. Hypothesis Testing. Similar to the standard logistic regression analysis, it is of interest to test for whether the regression coefficient (β) in model (2.2) is significantly different from zero or not. That is, the following hypotheses are tested for each region:

$$H_0 : \beta = 0 \ \ vs. \ \ H_1 : \beta \neq 0.$$

This corresponds to a test for independence, which determines whether there is a statistically significant relationship between the methylation level of a region and the condition of interest. If the null hypothesis is rejected, then the region is said to be differentially methylated. The Wald test statistic (Agresti, 1996) is used to conduct the test. A p-value is calculated for each region and adjusted p-values based on controlling the false discovery rate at α=0.05 across all regions are found (Benjamini and Hochberg, 1995). Any region with an adjusted p-value less than 0.05 is said to be differentially methylated.

2.4. FUZZY LOGISTIC REGRESSION

This section presents an alternative logistic regression approach based on fuzzy modeling for detecting differentially methylated regions (DMRs). The previous methods focused on different ways to summarize methylation levels over a region, which were considered as the response in an appropriate logistic regression model. A different way to view the region-level methylation is to consider that it has an element of vagueness, in that there is some ambiguity concerning the methylation level of the region. Viewed in this way, the region-level methylation outcome can be considered a fuzzy output. In this research, fuzzy logistic regression modeling (Pourahmad, 2013; Pourahmad *et al.*, 2011) techniques are explored to model the relationship between the vague region-level methylation and the factor of interest (e.g., disease vs. healthy). In this section, a review of fuzzy sets and fuzzy modeling is provided and a novel method for DMR testing using a fuzzy logistic regression approach is proposed. To the best of the author's knowledge, this method represents the first such application of fuzzy modeling to analyzing DNA methylation data. As such, several introductory concepts are presented for background of the proposed approach.

2.4.1. Introduction to Fuzzy Sets and Fuzzy Modeling.

Fuzzy set theory was first introduced by Lofti A. Zadeh in 1965 (Zadeh, 1965) and has since been used in many studies in medicine, biology, engineering, and other applications. In classical set theory, an element is either a member of the set or it is not. For example, consider the set of all even integers. All even integers are members of the set and all other possible values are not a member of the set. However, some situations exist where the set does not have such a precisely defined criteria for membership. For example, medical studies of patients with "high" blood pressure or methylation studies with regions of "low" methylation. This ambiguity in definition gives rise to the possibility of a fuzzy set, whose elements may have some degree of membership between 0 and 1. More formally, the following definitions of classical and fuzzy sets are taken from Pourahmad (2013).

Definition 1 (Classical Set): A classical set is defined as collection of elements of a universal ($A \subseteq X$) set that can be finite or countable. Each element can either belong to or not belong to a set ($x \in A$ or $x \notin A$). In the former case, the statement "x belongs to A" is true, whereas in the latter case this statement is false.

Definition 2 (Fuzzy Set): A fuzzy set of the universal set X is defined as a set of ordered pairs:

$$\tilde{A} = \{x, \mu_{\tilde{A}}(x) | x \in X\}$$

where, $\mu_{\tilde{A}}(.)$ is called the membership function of $\tilde{A}$, and $\mu_{\tilde{A}}(x)$ is the grade of membership of x in $\tilde{A}$. Note that $\mu_{\tilde{A}}(x)$ is also sometimes denoted $\tilde{A}(x)$.

A classical set is often referred to as a crisp set in this context and is a special case of when the membership function only takes on values 0 and 1. That is, the membership function of a classical set (A) is denoted as:

$$\mu_A(x) = \begin{cases} 1 & x \in A \\ 0 & x \notin A. \end{cases}$$

For fuzzy sets, values of the membership function range from 0 to 1, with 0 indicating non-membership, 1 indicating full membership and values in between indicating partial membership for the vague elements.

Fuzzy set theory is an extension of classical set theory that offers an alternative way to model uncertainty by providing a representation of observations and relationships that are vague or imprecise. Many additional concepts needed for fuzzy set theory, including fuzzy arithmetic, are given in Pourahmad (2013) and Zimmerman (1996). Only concepts needed for the formulation of the proposed method for DMR detection are provided in this dissertation. The following definitions that are needed for later formulations are provided from Pourahmad (2013).

Definition 3 (α-level Set): A set of elements that belong to a fuzzy set $\tilde{A}$ at least to the degree α is called α-level set of $\tilde{A}$:

$$A_\alpha = \{x \in X | \tilde{A}(x) \geq \alpha\}.$$

Definition 4 (Fuzzy Number): A fuzzy set of R for which any α-cut, $0 \leq \alpha \leq 1$, is a nonempty closed bounded interval is called a fuzzy number (denoted by E).

Definition 5 (LR-type, Triangular and Symmetric Triangular Fuzzy Numbers): A fuzzy number $\tilde{N} \in E$ is of LR-type (L stands for left and R stands for right) if it has the membership function as follows:

$$\tilde{N}(x) = \begin{cases} L(\frac{m-x}{\alpha}) & x \leq m \\ R(\frac{x-m}{\beta}) & x > m \end{cases}$$

where L and R are decreasing shape functions from R^+ to [0,1] with $L(0) = 1$, $L(x) < 1$ for all $x > 0$, $L(x) > 0$ for all $x < 1$ and $L(1) = 0$. Similar conditions hold for R. The real number m is called the mean value of $\tilde{N}$ and α and β (positive numbers) are called the left and right spreads, respectively. Symbolically $\tilde{N}$ is denoted by $(m, \alpha, \beta)_{LR}$. In the special case, where $L(x) = R(x) = max(0, 1 - x)$, then $\tilde{N}$ is called a triangular fuzzy number and denoted by $(m, \alpha, \beta)_T$. So, its membership function is:

$$\tilde{N}(x) = \begin{cases} 1 - \frac{m-x}{\alpha}, & m - \alpha \leq x \leq m \\ 1 - \frac{x-m}{\beta}, & m < x \leq m + \beta. \end{cases}$$

If, in addition, $\alpha = \beta$, then $\tilde{N}$ is denoted by $(m, \alpha)_T$ and is called the symmetric triangular fuzzy number.

Fuzzy set theory can be used in modeling the relationship between explanatory and response variables. In a conventional regression analysis, both the dependent (response) variable and random error terms are assumed to follow a probability distribution. However, many research studies involve vague or non-precise observations. When the response variable is a fuzzy observation, such distributional assumptions are not appropriate and a different type of modeling is needed. Tanaka *et al.* (1982) first proposed linear regression analysis with a fuzzy model as an extension of classical regression analysis. In such models, uncertainty is accounted for via fuzziness rather than including a random error term.

In fuzzy regression analysis, some elements of the model are fuzzy numbers (Pourahmad, 2013). There are different approaches to the fuzzy modeling depending on which elements of the model are fuzzy. Some models incorporate fuzziness in the relationships between variables, by incorporating fuzzy regression coefficients. The variables themselves may also be fuzzy numbers. In some models, only the response variable is fuzzy (and the explanatory variables are crisp), while in others both types of variables are fuzzy. Tanaka *et al.* (1982) first considered the case where the explanatory variables are crisp, the response is fuzzy and the relationship between the explanatory and response is fuzzy. Several other model formulations and estimation methods (Yen *et al.*, 1999) have since been proposed for fuzzy linear regression modeling.

In this work, fuzzy logistic regression (FLR) models are explored for situations when the response is a vague binary observation. Pourahmad (2013) proposed three different types of fuzzy logistic regression models and describe methods for estimation of the model parameters. A method for DMR detection based on FLR is developed based on these ideas. The model formulation and proposed approach are described in the following sections.

2.4.2. Theory of the Fuzzy Logistic Regression Model. As described previously, logistic regression is used in many fields to model the relationship between some explanatory variables and a categorical response variable. In a standard logistic regression analysis, the response is binary and extensions have been developed for multi-category and ordinal data. When the definition of the categories for a response variable are not precise, values may have a vague status and be viewed as a fuzzy output. For example, in the medical field consider the outcome of whether an individual has a certain disease or not. Some diseases do not have precisely established criteria for diagnosis and physicians may utilize information from multiple sources (e.g., symptoms, lab tests). Many patients may exhibit some but not all of the disease indicators and thus their true disease status remains vague. In addition, some ordinal responses (e.g., pain severity) that are described by linguistic terms such as low, medium, high can also be considered vague. These fuzzy categorical outputs can be modeled in a FLR framework.

The catalyst for this work is to develop improved methods for detecting differentially methylated regions using next-generation sequencing data. Testing for differential methylation for a region requires obtaining a region-level methylation value that represents a summary of the methylation levels at individual sites in the region. There are different ways to perform this region-level summary, which provide different information about the region (Table 2.1, Figure 2.1). An alternative is to consider the region-level methylation status as a vague or fuzzy binary or categorical observation.

Let the input (explanatory) and output (response) data consist of the following observations:

$$(x_{i1}, \ldots, x_{i(p-1)}, \tilde{Y}_i)$$

where x_{ij} are real crisp (explanatory or input) values in R for individual $1 \leq i \leq n$ and input variable $j = 1, \ldots, p-1$. Note that for DMR testing, one explanatory variable is of interest (e.g., treatment or disease status), but the model is given with additional predictors for thoroughness. $\tilde{Y}_i$ is a fuzzy (response or output) observation representing the methylation status of individual i for a particular region. The fuzzy observation ($\tilde{Y}_i$) can take two labels: approximately 1 or approximately 0, instead of 1 or 0 (Pourahmad, 2013). Due to the vague methylation status of the region, the binary response observations are not precise. This results in several issues for a standard logistic regression model. The Bernoulli probability distribution cannot be assumed and the probability of success ($P(Y_i = 1) = \pi_i$) cannot be calculated exactly. A model for π_i or the odds ($\frac{\pi_i}{1-\pi_i}$) based on explanatory variables is not meaningful in this situation. In this case, the possibility of success is considered instead of the probability of success. Possibility is another aspect of uncertainty. It measures the consistent degree of an individual's (methylation) status to a known or previously defined criteria (for methylation) which is called a success. A term called a *Possibilistic odds* is then defined and modeled. The following definition for is taken from Pourahmad (2013).

Definition 6 (Possibility and Possibilistic Odds): Let $\mu_i, i = 1, \ldots, n$, be the possibility of success, $\mu_i = poss(Y_i = 1)$, where μ_i is the consistent degree of the i^{th} individual to the success criteria. It can be defined in two ways:

- A real crisp value, $\mu_i \in R : 0 \leq \mu_i \leq 1$, or
- A linguistic term, $\mu_i \in \{\text{Very Low, Low, Medium, High, Very High}\}$. These logistic terms should be defined in such a way that the union of their support covers the whole range of (0, 1).

The ratio $(\frac{\mu_i}{1-\mu_i}), i = 1, \ldots, n$ is called the possibility odds of the i^{th} individual, which indicates the possibility of being methylated (success) status relative to the possibility of not being methylated (non success). See Pourahmad (2013) and Dubois and Prade (1988) for further details on possibility theory. To distinguish a crisp possibility from a linguistic one, the former will be denoted μ_i and the latter $\tilde{\mu}_i$.

When the possibility is a crisp value, the researcher compares each individual to criteria for a success and assigns a consistent degree as a real number in the (0,1) interval (Pourahmad, 2013). Alternatively, this work utilizes a linguistic term to define the possibility of success. These possibilities are defined by five different values for the linguistic term: $\tilde{\mu}_i \in \{\text{Very Low, Low, Medium, High, Very High}\}$. This approach involves assigning each individual a linguistic term based on the methylation data at sites on the region. First, k-means (k=5) clustering is performed and each site is assigned to one of values in $\tilde{\mu}_i$. Then a region-level value for $\tilde{\mu}_i$ is obtained by taking the mode of the linguistic terms across all sites in the region. This yields one linguistic term $\tilde{\mu}_i$ per region for each individual that indicates the possibility of being methylated.

To model the relationship between the crisp explanatory variables ($x'_{ij}s$) and the fuzzy responses ($\tilde{Y}_i$) with linguistic terms $\tilde{\mu}_i$, the following fuzzy logistic regression model is proposed:

$$\tilde{W}_i = ln(\frac{\tilde{\mu}_i}{1-\tilde{\mu}_i}) = \tilde{b}_0 + \tilde{b}_1 x_{i1} + \ldots + \tilde{b}_{p-1} x_{i(p-1)}, i = 1, \ldots, n. \qquad (2.3)$$

Here, $\tilde{b}_0, \tilde{b}_1, \ldots \tilde{b}_{p-1} \in E$ are fuzzy regression coefficients indicating a fuzzy relationship. One can transform the estimated outputs ($\tilde{W}_i$) into the possibility of success ($\tilde{\mu}_i(x)$) by using the extension principle (Wu, 2003). The definitions and propositions needed to achieve this goal are given below from Pourahmad (2013).

Definition 7 (Extension Principle): Let X be the Cartesian product of universes $X_1 \times X_2 \ldots \times X_n$ and $\tilde{A}_1, \tilde{A}_2, \ldots, \tilde{A}_n$ be n fuzzy sets in $X_1, \ldots, X_n$, respectively. Suppose that f is a mapping from X to a universe Y, $y = f(x_1, \ldots, x_n)$. Then, the extension principle allows a fuzzy set $\tilde{l}$ in Y to be defined by (Zimmerman, 1996):

$$\tilde{l} = \{(y, \tilde{l}(y)) | y = f(x_1, \ldots, x_n), (x_1, \ldots, x_n) \in X\}$$

where

$$\tilde{l}(y) = \begin{cases} \sup\limits_{(x_1,\ldots,x_n) \in f^{-1}} min\{\tilde{A}_1(x_1), \ldots, \tilde{A}_n(x_n)\} & if \quad f^{-1}(y) = 0 \\ 0 & otherwise \end{cases}$$

where f^{-1} is the inverse image of f.

In one dimension there is:

$$\tilde{l} = \{(y, \tilde{l}(y)) | y = f(x), x \in X\}$$

where

$$\tilde{l}(y) = \begin{cases} \sup\limits_{x \in f^{-1}} \tilde{A}(x) & if \quad f^{-1}(y) = 0 \\ 0 & otherwise. \end{cases}$$

The extension principle results in the following propositions (Pourahmad, 2013).

Proposition 1: Let $\tilde{M} = (m, \alpha, \beta)_{LR}$ be an LR-type fuzzy number and $c \in R$. Then

$$c\tilde{M} = \begin{cases} (cm, c\alpha, c\beta)_{LR} & c > 0 \\ (cm, -c\beta, -c\alpha)_{RL} & c < 0. \end{cases}$$

Proposition 2: Let $\tilde{M} = (m_1, \alpha_1, \beta_1)_{LR}$ and $\tilde{N} = (m_2, \alpha_2, \beta_2)_{LR}$ be two LR-type fuzzy numbers. Then,

$$\tilde{M} + \tilde{N} = (m_1 + m_2, \alpha_1 + \alpha_2, \beta_1 + \beta_2)_{LR}.$$

The observed outputs are given as, $\tilde{w}_i = ln(\frac{\tilde{\mu}_i}{1-\tilde{\mu}_i})$ for $i = 1, \ldots, n$. A membership function of these observed possibilistic odds $\tilde{w}_i$ can be derived once a formal definition of the $\tilde{\mu}_i$ are given. Suppose $\tilde{\mu}_i$ is the possibility of success defined by a linguistic variable with $u \in U = (0, 1)$. The values that this linguistic variable can take are:

$$T(x) = \{Very\ Low,\ Low,\ Medium,\ High,\ Very\ High\}.$$

Note that each of these values can be represented by a fuzzy set. Consider $\tilde{M}(x)$ to be a rule that assigns a fuzzy set to each value of the linguistic variable. In this work, the following definitions as proposed in Pourahmad (2013) are used:

$\tilde{M}(very\ low) = \{(u, \mu_{very\ low}(u)) | u \in (0,1)\}$ where

$$\mu_{very\ low}(u) = \begin{cases} 1 - \frac{0.02-u}{0.01} & 0.01 \le u \le 0.02 \\ 1 - \frac{u-0.02}{0.18} & 0.02 < u \le 0.18 \\ 0 & otherwise, \end{cases}$$

$\tilde{M}(low) = \{(u, \mu_{low}(u))|u \in (0,1)\}$ where

$$\mu_{low}(u) = \begin{cases} 1 - \frac{0.25-u}{0.15} & 0.1 \le u \le 0.25 \\ 1 - \frac{u-0.25}{0.15} & 0.25 < u \le 0.40 \\ 0 & otherwise, \end{cases}$$

$\tilde{M}(medium) = \{(u, \mu_{medium}(u))|u \in (0,1)\}$ where

$$\mu_{medium}(u) = \begin{cases} 1 - \frac{0.50-u}{0.15} & 0.35 \le u \le 0.5 \\ 1 - \frac{u-0.50}{0.15} & 0.5 < u \le 0.65 \\ 0 & otherwise, \end{cases}$$

$\tilde{M}(high) = \{(u, \mu_{high}(u))|u \in (0,1)\}$ where

$$\mu_{high}(u) = \begin{cases} 1 - \frac{0.75-u}{0.15} & 0.60 \le u \le 0.75 \\ 1 - \frac{u-0.75}{0.15} & 0.75 < u \le 0.90 \\ 0 & otherwise, \end{cases}$$

and $\tilde{M}(very\ high) = \{(u, \mu_{very\ high}(u))|u \in (0,1)\}$ where

$$\mu_{very\ high}(u) = \begin{cases} 1 - \frac{0.98-u}{0.18} & 0.80 \le u \le 0.98 \\ 1 - \frac{u-0.98}{0.01} & 0.98 < u \le 0.99 \\ 0 & otherwise. \end{cases}$$

The membership function of the observed outputs $(\tilde{w}_i)$ can then be calculated from the defined membership function of $\tilde{\mu}_i$ by the extension principle as follows:

$$f(x) = ln\frac{x}{1-x}, \quad 0 < x < 1$$

yields

$$\tilde{w}_i(y) = \sup_{\forall x: ln\frac{x}{1-x}=y} \tilde{\mu}_i(x).$$

Since $f(x) = ln\frac{x}{1-x}$. is a one-to-one function, there is only one $x \in (0, 1)$ such that $ln\frac{x}{1-x} = y$. Thus, the membership function of the logarithm transformation of possibility odds $\tilde{w}_i$ is derived as:

$$\tilde{w}_i(y = ln\frac{x}{1-x}) = \tilde{\mu}_i(\frac{exp(x)}{1+exp(x)}).$$

This result yields the following observed outputs for the fuzzy logistic regression model (Pourahmad, 2013).

$$\tilde{w}_{very\ low} = \begin{cases} 1 - \frac{0.02-(\frac{exp(x)}{1+exp(x)})}{0.01} & -4.59 \leq x \leq -3.89 \\ 1 - \frac{(\frac{exp(x)}{1+exp(x)})-0.02}{0.18} & -3.89 < x \leq -1.38 \\ 0 & otherwise, \end{cases}$$

$$\tilde{w}_{low} = \begin{cases} 1 - \frac{0.25-(\frac{exp(x)}{1+exp(x)})}{0.15} & -2.19 \leq x \leq -1.09 \\ 1 - \frac{(\frac{exp(x)}{1+exp(x)})-0.25}{0.15} & -1.09 < x \leq -0.40 \\ 0 & otherwise, \end{cases}$$

$$\tilde{w}_{medium} = \begin{cases} 1 - \frac{0.50-(\frac{exp(x)}{1+exp(x)})}{0.15} & -0.62 \leq x \leq 0 \\ 1 - \frac{(\frac{exp(x)}{1+exp(x)})-0.50}{0.15} & 0 < x \leq 0.62 \\ 0 & otherwise, \end{cases}$$

$$\tilde{w}_{high} = \begin{cases} 1 - \frac{0.75-(\frac{exp(x)}{1+exp(x)})}{0.15} & 0.41 \leq x \leq 1.099 \\ 1 - \frac{(\frac{exp(x)}{1+exp(x)})-0.75}{0.15} & 1.099 < x \leq 2.20 \\ 0 & otherwise, \end{cases}$$

and

$$\tilde{w}_{very\ high} = \begin{cases} 1 - \frac{0.98-(\frac{exp(x)}{1+exp(x)})}{0.18} & 1.39 \leq x \leq 3.90 \\ 1 - \frac{(\frac{exp(x)}{1+exp(x)})-0.98}{0.01} & 3.90 < x \leq 4.60 \\ 0 & otherwise. \end{cases}$$

2.4.3. Fuzzy Least Squares Method for Estimating FLR Parameters. The FLR model (2.3) has a set of fuzzy regression coefficients $(\tilde{b}_0, \tilde{b}_1, \ldots \tilde{b}_{p-1})$ that require estimation. This research utilizes an extension of the ordinary least squares approach to fuzzy regression modeling for fuzzy response data in which the parameters of the model are also assumed to be fuzzy numbers. This fuzzy least squares method is detailed in Pourahmad (2013) and an explanation is provided below for thoroughness. It assumes that all fuzzy data are symmetric LR fuzzy numbers, which is proposed by Diamond (1987) and Celmiņš (1987). First, a definition for the distance in the space between fuzzy numbers is required (Pourahmad, 2013).

Definition 8 (Distance Between Two Fuzzy Numbers): Let $\tilde{A}, \tilde{B} \in E$ be two fuzzy numbers. Also, suppose $\tilde{A}_\alpha = [a_1, a_2]$ and $\tilde{B}_\alpha = [b_1, b_2]$ are their respective α-cuts. The distance between these numbers is defined as follows:

$$D(\tilde{A}, \tilde{B}) = \left[\int_0^1 f(\alpha) d^2(\tilde{A}_\alpha, \tilde{B}_\alpha) d\alpha \right]^{\frac{1}{2}}$$

where

$$d^2(\tilde{A}_\alpha, \tilde{B}_\alpha) = [a_1 - b_1]^2 + [a_2 - b_2]^2,$$

and $f(\alpha)$ is an increasing function of $[0, 1]$ for which $f(0) = 0$ and $\int_0^1 f(\alpha) d\alpha = \frac{1}{2}$. The function $f(\alpha)$ is a weighting function for $d^2(\tilde{A}_\alpha, \tilde{B}_\alpha)$. In determining the distance between two fuzzy numbers $(\tilde{A}, \tilde{B})$, the distance function $(D(\tilde{A}, \tilde{B}))$ is a measure of discrepancy between $\tilde{A}$ and $\tilde{B}$. Using ideas from ordinary least squares, the goal is to obtain estimates $(\tilde{b}_0, \tilde{b}_1, \ldots \tilde{b}_{p-1})$ such that the distance between the observed and estimated values is minimized. For the proposed model, this corresponds to minimizing the sum of the distance between $\tilde{w}_i$ and $\hat{\tilde{W}}_i$, which is called the sum of squared error (SSE). The SSE is defined as follows:

$$SSE = \sum_{i=1}^{n} \left(d(\tilde{w}_i, \tilde{W}_i) \right)^2 .$$

To estimate the fuzzy parameters, the partial derivatives of the SSE with respect to $\tilde{b}_j, j = 0, 1, \ldots, p-1$ are set to 0 and solved:

$$\frac{\partial}{\partial \tilde{b}_j} SSE = 0, j = 0, 1, \ldots, p-1.$$

Without loss of generality, assume that $\tilde{b}_j = (a_j, s_j)$ are symmetric triangular fuzzy numbers. Then, the estimated outputs are also symmetric triangular fuzzy numbers:

$$(\tilde{W}_i)_\alpha = [(\alpha - 1) f_i(s) + f_i(a), (1 - \alpha) f_i(s) + f_i(a)]$$

in which $f_i(a) = a_0 + a_1 x_{i1} + \ldots + a_n x_{in}$, and $f_i(s) = s_0 + s_1 x_{i1} + \ldots + s_n x_{in}$ by some algebraic calculation. To calculate $(\tilde{w}_i)_\alpha$ based on $(\tilde{\mu}_i)_\alpha$, there is $\tilde{\mu}_i = (m_i, d_i^L, d_i^R)$ and

$$(\tilde{\mu}_i)_\alpha = [(\alpha - 1) d_i^L + m_i, (1 - \alpha) d_i^R + m_i];$$

therefore

$$(\tilde{w}_i)_\alpha = \left[ln \frac{(\alpha - 1) d_i^L + m_i}{1 - (\alpha - 1) d_i^L - m_i}, ln \frac{(1 - \alpha) d_i^R + m_i}{1 - (1 - \alpha) d_i^R - m_i} \right].$$

To calculate the distance between $\tilde{w}_i$ and $\tilde{W}_i$:

$$d(\tilde{w}_i, \tilde{W}_i) = \left[\int_0^1 g(\alpha) d^2((\tilde{w}_i)_\alpha, (\tilde{W}_i)_\alpha) d\alpha \right]^{\frac{1}{2}}$$

where $g(\alpha) = \alpha$ is the weighting function, $\int_0^1 g(\alpha)d\alpha = \frac{1}{2}$, $g(0) = 0$ and $d^2((\tilde{w}_i)_\alpha, (\tilde{W}_i)_\alpha) = \left[ln\frac{(\alpha-1)d_i^L+m_i}{1-(\alpha-1)d_i^L-m_i} - (\alpha-1)f_i(s) - f_i(a)\right]^2 + \left[ln\frac{(1-\alpha)d_i^R+m_i}{1-(1-\alpha)d_i^R-m_i} - (1-\alpha)f_i(s) - f_i(a)\right]^2.$

The SSE criterion that should be minimized when estimating the parameters is:

$$SSE = \sum_{i=1}^{n}\left(\int_0^1 g(\alpha)\big[\text{fuzzy number left} + \text{fuzzy number right}\big]d\alpha\big]\right)$$

where

$$\text{fuzzy number left} = \left[ln\frac{(\alpha-1)d_i^L + m_i}{1-(\alpha-1)d_i^L - m_i} - (\alpha-1)f_i(s) - f_i(a)\right]^2$$

and

$$\text{fuzzy number right} = \left[ln\frac{(1-\alpha)d_i^R + m_i}{1-(1-\alpha)d_i^R - m_i} - (1-\alpha)f_i(s) - f_i(a)\right]^2.$$

The distance function depends on the model's coefficients only through $f_i(s)$ and $f_i(a)$. Partial derivatives of the SSE with respect to a_j and s_j equal to zero as follows:

$$\frac{\partial}{\partial a_j}SSE = 0 \quad and \quad \frac{\partial}{\partial s_j}SSE = 0.$$

This leads to the following equations:

$$\sum_{i=1}^{n}\left(\int_0^1 2\alpha x_{ij}\left[2f_i(a) - ln\frac{(\alpha-1)d_i^L + m_i}{1-(\alpha-1)d_i^L - m_i} - ln\frac{(1-\alpha)d_i^R + m_i}{1-(1-\alpha)d_i^R - m_i}\right]d\alpha\right) = 0 \quad (2.4)$$

and

$$\sum_{i=1}^{n}\left(\int_0^1 2\alpha(1-\alpha)x_{ij}\left[2(1-\alpha)f_i(s) + ln\frac{(\alpha-1)d_i^L + m_i}{1-(\alpha-1)d_i^L - m_i} - ln\frac{(1-\alpha)d_i^R + m_i}{1-(1-\alpha)d_i^R - m_i}\right]d\alpha\right) = 0. \quad (2.5)$$

The quantities in the above equations depend on $(\tilde{\mu}_i)$ assigned to each individual. For DNA methylation data, this is the mode of the linguistic values a cross the region. From equation 2.4 the following is obtained.

$$\sum_{i=1}^{n}\left(\int_0^1 \left(4\alpha x_{ij} f_i(a) d\alpha\right) - \int_0^1 \left(2\alpha x_{ij} ln \frac{(\alpha-1)d_i^L + m_i}{1-(\alpha-1)d_i^L - m_i}\right)\right.$$

$$\left. - \int_0^1 \left(2\alpha x_{ij} ln \frac{(1-\alpha)d_i^R + m_i}{1-(1-\alpha)d_i^R - m_i} d\alpha\right)\right) = 0$$

and

$$\sum_{i=1}^{n}\left(2x_{ij} f_i(a) - 2x_{ij} z_{i1} - 2x_{ij} z_{i2}\right) = 0$$

where $z_{i1} = \int_0^1 \left(\alpha ln \frac{(\alpha-1)d_i^L+m_i}{1-(\alpha-1)d_i^L-m_i} d\alpha\right)$ and $z_{i2} = \int_0^1 \left(\alpha ln \frac{(1-\alpha)d_i^R+m_i}{1-(1-\alpha)d_i^R-m_i} d\alpha\right)$ are two real numbers. From equation 2.5, the following is obtained:

$$\sum_{i=1}^{n}\left(\int_0^1 4\alpha(1-\alpha)^2 x_{ij} f_i(s) d\alpha + \int_0^1 2\alpha(1-\alpha) x_{ij} ln \frac{(\alpha-1)d_i^L + m_i}{1-(\alpha-1)d_i^L - m_i} d\alpha\right.$$

$$\left. - \int_0^1 2\alpha(1-\alpha) x_{ij} ln \frac{(1-\alpha)d_i^R + m_i}{1-(1-\alpha)d_i^R - m_i} d\alpha\right) = 0$$

and

$$\sum_{i=1}^{n}\left(\frac{1}{3} x_{ij} f_i(s) + 2x_{ij} k_{i1} - 2x_{ij} k_{i2}\right) = 0$$

where $k_{i1} = \int_0^1 \left(\alpha(1-\alpha) ln \frac{(\alpha-1)d_i^L+m_i}{1-(\alpha-1)d_i^L-m_i} d\alpha\right)$ and $k_{i2} = \int_0^1 \left(\alpha(1-\alpha) ln \frac{(1-\alpha)d_i^R+m_i}{1-(1-\alpha)d_i^R-m_i} d\alpha\right)$ are two real numbers. Finally, the following equations are derived:

$$a_0 \sum_{i=1}^{n} x_{i0} x_{ij} + a_1 \sum_{i=1}^{n} x_{i1} x_{ij} + \ldots + a_n \sum_{i=1}^{n} x_{in} x_{ij} = \sum_{i=1}^{n} z_i x_{ij} \tag{2.6}$$

and

$$s_0 \sum_{i=1}^{n} x_{i0} x_{ij} + s_1 \sum_{i=1}^{n} x_{i1} x_{ij} + \ldots + s_n \sum_{i=1}^{n} x_{in} x_{ij} = \sum_{i=1}^{n} k_i x_{ij} \tag{2.7}$$

where $x_{i0} = 1$ and z_i ,k_i are the results of integral computation in the estimation process for each case. Equations 2.6 and 2.7 can be written in matrix form as follows. First, $Aa = Z$, where $A = X'X$,

$$X = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1(p-1)} \\ 1 & x_{21} & \cdots & x_{2(p-1)} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{n(p-1)} \end{bmatrix}$$

is the design matrix, $a = (a_0, a_1, \ldots, a_{p-1})^T$, and

$$Z = \left(\sum_{i=1}^{n} z_i x_{i0}, \sum_{i=1}^{n} z_i x_{i1}, \ldots, \sum_{i=1}^{n} z_i x_{i(p-1)} \right)^T .$$

Also in matrix form, $As = K$ where $s = (s_0, s_1, \ldots, s_{p-1})^T$, and

$$K = \left(\sum_{i=1}^{n} k_i x_{i0}, \sum_{i=1}^{n} k_i x_{i1}, \ldots, \sum_{i=1}^{n} k_i x_{i(p-1)} \right)^T .$$

Thus, the estimates $(\hat{\tilde{b}}_j)$ for $\tilde{b}_j = (a_j, s_j)$ are: $a = A^{-1}Z$ and $s = A^{-1}K$. Note that a separate model is fit for each region.

2.4.4. Hypothesis Testing for Fuzzy Regression Coefficients. Although Pourahmad (2013) provided the inspiration for using fuzzy logistic regression for DMR testing, that approach does not provide a method for hypothesis testing of the fuzzy regression coefficients. A bootstrap-based approach is proposed by Lee *et al.* (2015) for fuzzy linear regression to conduct hypothesis tests for the fuzzy regression coefficients. This work employs the same type of bootstrap approach to determine the significance of the fuzzy coefficients in a fuzzy logistic regression model. Testing for significance of the fuzzy coefficient corresponding to the factor of interest provides a test for differential methylation at the region-level. Consider the hypotheses as follows:

$$H_0 : \tilde{b}_{j\alpha} = \{0\} \quad \text{vs.} \quad H_1 : \tilde{b}_{j\alpha} \neq \{0\},$$

where $j = 0, 1, \ldots, p - 1$. The test statistic proposed by Lee *et al.* (2015) is used as follows:

$$T_{j\alpha} = \sqrt{\frac{d^2(\hat{\tilde{b}}_{j\alpha}\Theta_g\{0\})}{S_{\hat{\tilde{b}}_{j\alpha}}}}$$

where $S_{\hat{\tilde{b}}_{j\alpha}} = \sqrt{g[j+1]\frac{\sum_{i=1}^{n} d^2(\tilde{w}_{i\alpha},\tilde{W}_{i\alpha})}{n-2}}$ and $g[j]$ is the j^{th} diagonal entry of the matrix $(X^T X)^{-1}$. The distance between the α-cut of two fuzzy numbers $\tilde{A}$ and $\tilde{B}$ is defined as follows: $d^2(\tilde{A}_\alpha, \tilde{B}_\alpha) = (\tilde{A}^C_\alpha - \tilde{B}^C_\alpha) + (\tilde{A}^W_\alpha - \tilde{B}^W_\alpha)$ where $\tilde{A}^C_\alpha = \frac{1}{2}(\tilde{A}^L_\alpha + \tilde{A}^U_\alpha)$ and $\tilde{A}^W_\alpha = \frac{1}{2}(\tilde{A}^L_\alpha - \tilde{B}^U_\alpha)$. The general Hukuhara difference is defined as $[\tilde{A}^L_\alpha, \tilde{A}^U_\alpha]\Theta_g[\tilde{B}^L_\alpha, \tilde{B}^U_\alpha] = [\tilde{C}^-_\alpha, \tilde{C}^+_\alpha]$ where $\tilde{C}^-_\alpha = min\{\tilde{A}^L_\alpha - \tilde{B}^L_\alpha, \tilde{A}^U_\alpha - \tilde{B}^U_\alpha\}$ and $\tilde{C}^+_\alpha = max\{\tilde{A}^L_\alpha - \tilde{B}^L_\alpha, \tilde{A}^U_\alpha - \tilde{B}^U_\alpha\}$. The p-value of the test can be computed as the proportion of values $T^1_{j\alpha}, \ldots, T^B_{j\alpha}$ that is greater than or equal to the calculated test statistic $T_{j\alpha}$:

$$p\text{-value}_\alpha = \frac{\#\{b : T^b_{j\alpha} \geq T_{j\alpha}\}}{B}$$

where B = the number of bootstrap samples. A separate test is conducted for each region and the false discovery rate was controlled across all the regions at 5% (Benjamini and Hochberg, 1995). Regions with an adjusted p-value < 0.05 are considered DMRs.

2.4.5. Summary of FLR Approach for Detecting DMRs. Fuzzy logistic regression (FLR) is an extension of classical logistic regression analysis in which some elements of the model are represented by fuzzy numbers. In this work, a FLR approach was developed to detect differentially methylation regions (DMRs). This approach provides an alternative way to view the methylation levels of a region as fuzzy observations that are vague or imprecise. These fuzzy observations are incorporated into a fuzzy modeling approach based on the FLR methodology in Pourahmad (2013) and bootstrap based hypothesis testing for fuzzy regression coefficients in Lee *et al.* (2015). The first step in the FLR approach for DMR detection involves performing k-means clustering (k=5) on the DNA methylation levels of all cytosine sites. Cutoffs based on the clustering results are then used to classify each cytosine site into methylation level groups (Very Low, Low, Medium, High and Very High). The mode across the region is used to identify a linguistic term for the region, which is considered the fuzzy response in the FLR model (2.3). The crisp explanatory variable is the indicator variable for the condition of interest (e.g., disease vs. healthy), although other covariates could be included. The fuzzy coefficients for the FLR model (2.3) relating the condition of interest to the fuzzy methylation output are then estimated via the fuzzy least squares method. A bootstrap-based approach is then used to test whether the fuzzy coefficient corresponding to the condition term is zero or not and p-value is calculated. Finally, the p-values are adjusted to control the false discovery rate (FDR) at 0.05 across all regions and a list of significant DMRs is obtained.

3. SIMULATION STUDIES

3.1. INTRODUCTION OF SIMULATION STUDIES

To evaluate the performance of the proposed models described in Section 2 for differentially methylated region (DMR) testing, two different simulation studies were conducted and are presented in this section. Both studies compare the true positive and false positive rates between the existing method, methylation analysis using genome information (MAGIg) (Baumann and Doerge, 2014), and the three proposed logistic regression methods: weighted average logistic regression (WALR), ordinal logistic regression (OLR), and fuzzy logistic regression (FLR). Both studies are based on real data to help create a data set that maintains a realistic structure. The first simulation study presented is based on plant data, which involves analyzing cytosine sites in the CG, CHG, and CHH contexts. Since the mechanism of methylation differs between these sequence contexts, it is important to evaluate the methods in each context separately. The second simulation study is based on human data, which only involves analyzing CG sites. Various settings are investigated in both simulations to determine different situations that may affect method performance.

3.2. SIMULATION STUDY I: PLANT DATA

To evaluate the performance of the four methods (MAGIg, WALR, OLR, FLR), a simulation study based on reduced representation bisulfite sequencing (RRBS) data was performed. Methylation data for three replicates of Columbia-0 *Arabidopsis thaliana* seedlings (Law *et al.*, 2013; Qian *et al.*, 2012; Zhong *et al.*, 2012) were accessed from the NGSmethDB website (https://bioinfo2.ugr.es/NGSmethDB/). *Arabidopsis thaliana* (or thale cress) is a small flowering plant that is often used as a model organism in the study of plants. Columbia-0 is a widely studied wild type accession of *Arabidopsis*. NGSmethDB is an online data repository for DNA methylation NGS data. Data are available on the cytosine context (CG, CHG, CHH), genomic location (chromosome, position), and the number of methylated (and total) reads for each cytosine. NGSmethDB also provides filtered data sets based on the sequencing depth. Sites with a minimum depth of 10 reads were chosen for further analysis.

3.2.1. Simulation Framework. To simulate methylation profiles in a realistic way, an approach similar to M3D (Mayo *et al.*, 2015) and Baumann *et al.* (2015) was taken using the RRBS data set described above. The simulation study focuses on the first 1000 regions (genes) on chromosome 1. The three replicates serve as the control group. A case group is simulated by applying some random noise to the three replicates and a subset of genes are simulated to be differentially methylated. First, following Baumann *et al.* (2015), the number of reads at each site is simulated by adding or subtracting random Poisson ($\delta = 1$) noise to the control samples. Then, random noise from a Uniform (-0.1, 0.1) distribution is added to the cytosine methylation level, L_i, which is the proportion of methylated reads out of the total number of reads for individual i at a particular site.

To simulate methylation changes, 200 genes were randomly selected out of a possible 1000 to be true DMRs and differential methylation changes were applied according to the simulation framework of Mayo *et al.* (2015). The methylation level L_i^{old} represents the proportion of methylated cytosines in control sample i at a particular site. The methylation level in the case group is represented by L_i^{new} and is simulated as follows. For hypermethylation (methylation higher in case than control), $L_i^{new} = (1 - \lambda)\, L_i^{old} + \lambda$ when $L_i^{old} \leq 0.5$ on average for the gene. For hypomethyation (methylation lower in case than control), $L_i^{new} = (1 - \lambda)\, L_i^{old}$ when $L_i^{old} > 0.5$ on average for the gene. The degree methylation change is controlled by the parameter $\lambda \in [0, 1]$.

To investigate the performance of the methods under different degrees of differential methylation, the λ parameter is varied as $\lambda = \{0.2, 0.3, 0.4, 0.5, 0.6\}$. Methylation changes were applied to all cytosines of a specific context (CG, CHG, CHH) within the selected genes that were DMRs. MAGIg, WALR, OLR, and FLR were applied for DMR testing separately for each of the three sequence contexts and across the various settings of λ. For FLR, there

Table 3.1. Results for CG sites. True positive rate and false positive rate are given for FLR, OLR, WALR, and MAGIg with various degrees of methylation change λ.

CG			
λ	Method	TPR	FPR
0.2	FLR	0.98	0.0025
0.2	OLR	0.95	0
0.2	WALR	0.77	0
0.2	MAGIg	0.60	0
0.3	FLR	0.99	0.0026
0.3	OLR	0.96	0
0.3	WALR	0.78	0
0.3	MAGIg	0.77	0
0.4	FLR	0.99	0.0025
0.4	OLR	0.96	0
0.4	WALR	0.78	0
0.4	MAGIg	0.77	0
0.5	FLR	0.99	0.0025
0.5	OLR	0.96	0
0.5	WALR	0.78	0
0.5	MAGIg	0.77	0
0.6	FLR	0.99	0.0026
0.6	OLR	0.97	0
0.6	WALR	0.79	0
0.6	MAGIg	0.78	0

are B=1000 bootstrap samples used in the test of the fuzzy regression coefficient. The false discovery rate (FDR) is controlled at 5% for each analysis. The methods are compared via the true positive rate (TPR) and false positive rate (FPR). The TPR is the proportion of true DMRs that are correctly detected. The TPR is also referred to as the observed power or sensitivity, with values closer to 1 indicating better performance. The FPR is the proportion of genes that are falsely identified as DMRs among the genes that are truly not differentially methylated. The FPR is the observed type I error rate, with values closer 0 (or the significance level α) being desired.

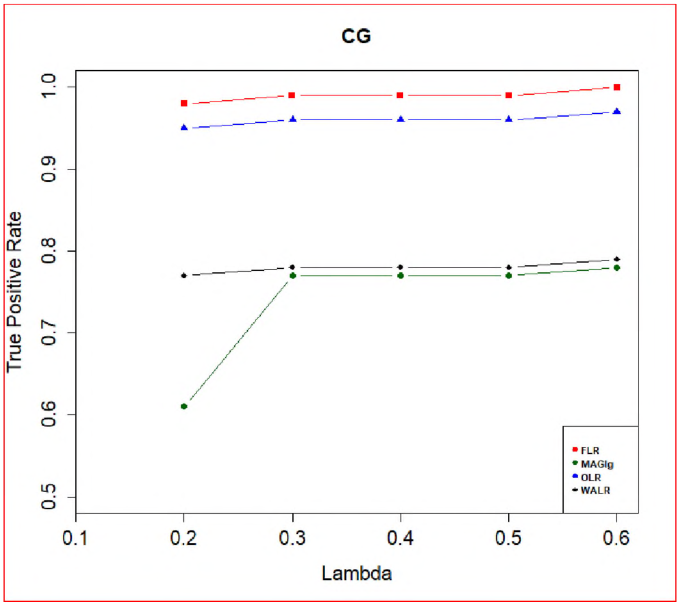

Figure 3.1. For CG sites, the true positive rate versus λ level for controlling the degree of differential methylation for each of the four methods (FLR-Red, MAGIg-Green, OLR-Blue, and WALR-Black).

3.2.2. Results. In the CG context, results for the four methods (FLR, OLR, WALR, and MAGIg) are summarized in Table 3.1 for different values of the methylation difference strength parameter λ. The TPR and FPR are given for each method. As an illustration of the calculations, the FLR method identified 196 out of 200 true DMRs, which yields a TPR of 0.98 when λ=0.2. Only 2 genes were falsely identified as DMRs out of a total of 800 genes that are truly non-DMRs, yielding a FPR of 0.0025 when λ=0.2. The FPR is zero for all of the methods except for FLR, where it is still small and controlled to be below 5%. The TPR for varying degrees of methylation difference λ are displayed in graphical form in Figure 3.1 for each of the four methods. The FLR method had the highest TPR across all λ values, with OLR yielding similar, but slightly lower, TPR. Both methods had high TPR values of 0.95 or higher for all λ values. The WALR and MAGIg methods did not perform as well. Both methods had TPR values around 0.77-0.79 at all λ values except when λ=0.2, in which MAGIg exhibited reduced performance (TPR=0.6).

Table 3.2. Results for CHG sites. True positive rate and false positive rate are given for FLR, OLR, WALR, and MAGIg with various degrees of methylation change λ.

CHG			
λ	Method	TPR	FPR
0.2	FLR	0.97	0.0025
0.2	OLR	0.95	0
0.2	WALR	0.88	0
0.2	MAGIg	0.73	0
0.3	FLR	0.98	0.0037
0.3	OLR	0.95	0
0.3	WALR	0.88	0
0.3	MAGIg	0.86	0
0.4	FLR	0.99	0.0025
0.4	OLR	0.96	0
0.4	WALR	0.88	0
0.4	MAGIg	0.86	0
0.5	FLR	0.99	0.005
0.5	OLR	0.96	0
0.5	WALR	0.88	0
0.5	MAGIg	0.86	0
0.6	FLR	0.99	0.0012
0.6	OLR	0.96	0
0.6	WALR	0.88	0
0.6	MAGIg	0.86	0

Results for the CHG context are summarized in Table 3.2 and Figure 3.2. These results are similar to the patterns observed in the CG context. The FPR is small (less than 0.05) for all methods and is zero for all methods except FLR. The FLR and OLR performed well in terms of TPR (0.95 or greater across all λ values), with FLR always slightly outperforming OLR. The WALR and MAGIg had reduced performance, but TPR values were slightly higher than in the CG context (between 0.86-0.88 for most values of λ). Again, MAGIg did not perform as well when the methylation difference was small (λ=0.2).

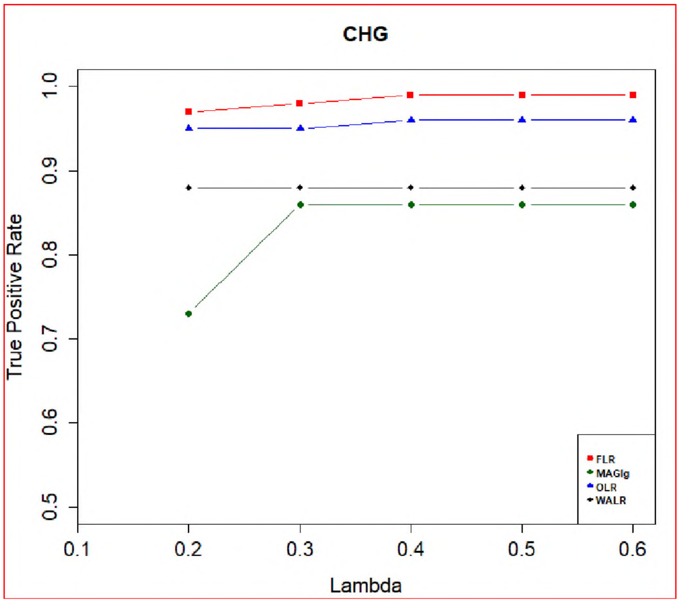

Figure 3.2. For CHG sites, the true positive rate versus λ level for controlling the degree of differential methylation for each of the four methods (FLR-Red, MAGIg-Green, OLR-Blue, and WALR-Black).

For the CHH context, the results are given in Table 3.3 and Figure 3.3. For the FPR, the results are similar to the CG and CHG context, with all methods being below 0.05. For the TPR, the results are more varied than they were for the other contexts. The TPR was at least 0.90 for all methods at λ values 0.3 and higher, indicating they all perform well. However there is not one method that is clearly superior across all values of λ. The methods all have a slightly lower TPR at λ=0.2, with MAGIg having the lowest TPR of 0.86.

In conclusion, all methods had a low FPR (<0.05) across all λ values and sequence contexts. All methods except the FLR actually had no false positives across all settings, but FLR still only had a very small proportion of false positives, well below 0.05. The TPR results for CG and CHG were similar and indicated a clear performance advantage of FLR and OLR, with FLR having a slightly better TPR across all settings. The results for CHH

Table 3.3. Results for CHH sites. True positive rate and false positive rate are given for FLR, OLR, WALR, and MAGIg with various degrees of methylation change λ.

CHH			
λ	Method	TPR	FPR
0.2	FLR	0.95	0.0025
0.2	OLR	0.92	0
0.2	WALR	0.90	0
0.2	MAGIg	0.86	0
0.3	FLR	0.96	0.0050
0.3	OLR	0.93	0
0.3	WALR	0.98	0
0.3	MAGIg	0.95	0
0.4	FLR	0.98	0.0025
0.4	OLR	0.95	0
0.4	WALR	0.98	0
0.4	MAGIg	0.95	0
0.5	FLR	0.98	0.0024
0.5	OLR	0.95	0
0.5	WALR	0.97	0
0.5	MAGIg	0.95	0
0.6	FLR	0.98	0.0022
0.6	OLR	0.95	0
0.6	WALR	0.97	0
0.6	MAGIg	0.95	0

were mixed with all methods performing fairly well. Although the TPR was similar across all λ values for a particular method and sequence context, the TPR was always lower when λ=0.2, which is expected since it is more difficult to detect smaller differences.

3.2.3. Correlation Levels for Different Contexts. One possible explanation for the performance differences between sequence contexts could be due to the differences in correlation between methylation levels of neighboring cytosines for the different contexts. To investigate this, 100 reference cytosine sites for each context were randomly select from the genes that were included in the simulation study. The Pearson correlation between methylation level of the reference and each cytosine up to 10 sites downstream of the

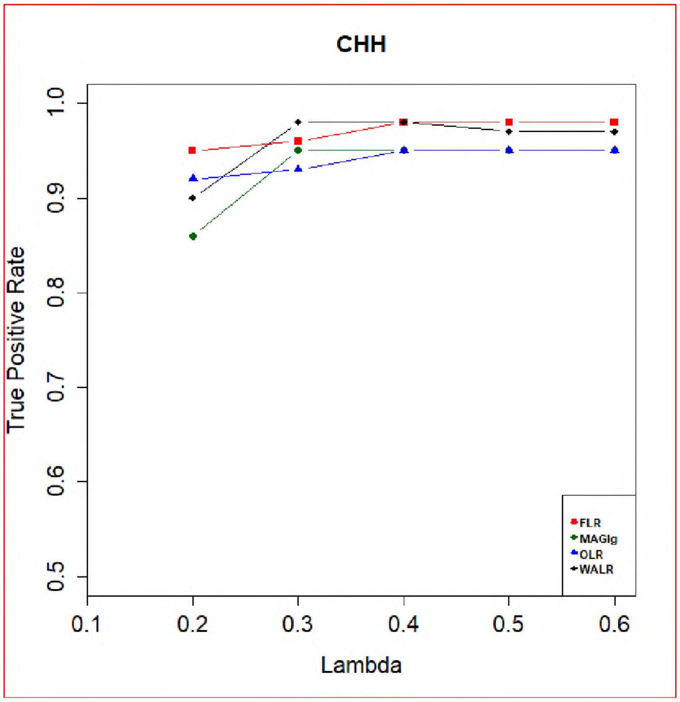

Figure 3.3. For CHH sites, the true positive rate versus λ level for controlling the degree of differential methylation for each of the four methods (FLR-Red, MAGIg-Green, OLR-Blue, and WALR-Black).

reference was calculated. Figure 3.4 shows the average Pearson correlation over the 100 locations between the reference and downstream cytosines under different sequence contexts. Three different coverage depths (1, 5, and 10 reads) were also investigated. In all contexts, the correlation between the methylation levels declines the further away a site is from the reference cytosine. CG tends to have the highest level of correlation between neighboring sites, followed closely by CHG. The correlation is greater than 0.8 up to 10 cytosines away from the reference in both contexts. The correlation is lower in CHH ranging from around 0.6 for the sites furthest away from the reference to 0.8 at sites closest to the reference. The correlation values are similar for all of the sequencing depths and there is not a clear trend when summarizing across contexts and number of sites away from the reference.

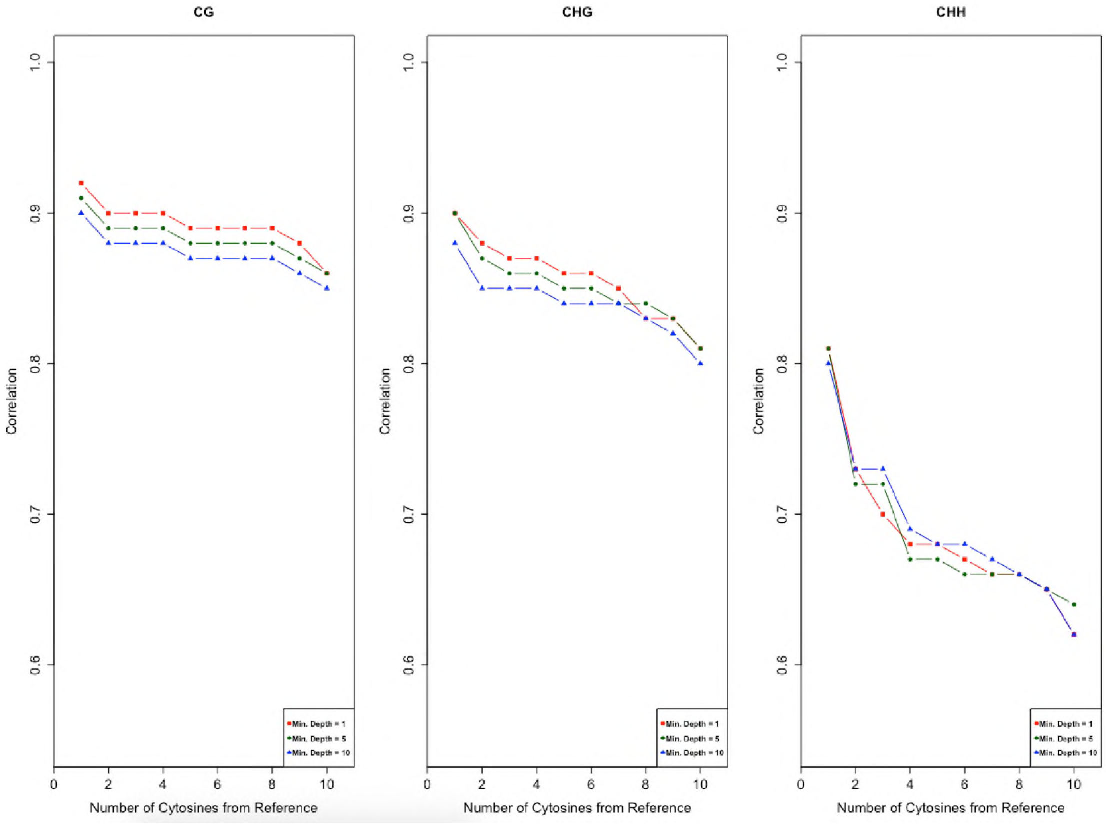

Figure 3.4. The average Pearson correlation of methylation levels between a reference cytosine and 10 downstream cytosines for 100 reference sites. Results are given for different cytosine contexts (CG, CHG, and CHH) and three coverage depths (1, 5 and 10 reads).

When viewed in light of the simulation study results, the TPR of FLR and OLR stayed fairly consistent across all contexts, whereas the TPR of WALR and MAGIg differed by sequence context. WALR and MAGIg had the highest TPR for the CHH context followed by the CHG context and was the worst for the CG context. It is interesting to note that performance improved for these two methods in contexts where the correlation with neighboring sites was lower (this is most noticeable for CHH). This is an observation that merits further investigation in future studies since it may offer insights into situations when WALR and MAGIg may be preferred. However, due to the consistent performance of FLR and OLR, these methods are recommended based on this simulation study with FLR having a slight advantage over OLR.

3.3. SIMULATION STUDY II: HUMAN DATA

To evaluate the performance of the four logistic regression methods (MAGIg, WALR, OLR, FLR) on human data and investigate the impact of the replicate number, a simulation study based on real RRBS data was performed. Methylation data of bisulfite-sequenced DNA were obtained from 4 patients with acute promyelocytic leukemia (APL) and 61 APL control samples. This data set was obtained under accession number GSE27584 from the Gene Expression Omnibus at the National Center for Biotechnology Information website (https://www.ncbi.nlm.nih.gov/geo). The RRBS data were pre-processed using Bismark version 0.5 (a reference genome alignment tool) that maps bisulfite-treated sequencing reads to a genome of interest and performs methylation calls in a single step. Only the CG context was available for analysis, since that is the primary methylation context in humans. In this simulation study, regions are defined based on CG density rather than genomic annotation (genes).

3.3.1. Simulation Framework. To realistically mimic methylation profile changes, a simulation was constructed from the RRBS data set described above following a similar approach as described for the plant data. The primary difference is the definition of the region, which follows the approach as in M3D (Mayo *et al.*, 2015). The regions (CG clusters) are defined as follows: (1) CG sites that cover at least 75% of samples are defined as frequently covered CG sites, and (2) CG sites with a maximum distance of 100 base pairs to the nearest neighbor are included in the region. Using these criteria, only regions with at least 20 frequently covered CG sites are used in the analysis.

The simulation study focuses on the first 1,000 regions of chromosome 1. Only a subset of 12 out of 61 control samples are used in simulation study. These are split into three different simulations of three, four, and five replicates to investigate the impact of sample size on the method performance. A case group is simulated for each of the different replicate numbers by first adding or subtracting random Poisson ($\delta = 1$) noise to the total number of reads at all cytosines in each control sample. Then, random noise from a Uniform (-0.1, 0.1) distribution is added to the cytosine methylation level, L_i, defined as the ratio of methylated reads to the total number of reads at a particular cytosine for individual i. Of the 1000 CG clusters (predefined regions), 300 are randomly selected for the application of differential methylation changes to be true DMRs. The methylation levels are adjusted within the 300 selected regions and the degree of methylation level change is controlled by the parameter $\lambda \in [0, 1]$. The methylation level L_i^{old} represents the proportion of methylated cytosines in control sample i at a particular site. New methylation levels for true DMRs in the case group were simulated by $L_i^{new} = (1 - \lambda)\ L_i^{old} + \lambda$ when $L_i^{old} \leq 0.5$ on average for hypermethylation and $L_i^{new} = (1 - \lambda)\ L_i^{old}$ when $L_i^{old} > 0.5$ on average for hypomethylation.

FLR, OLR, WALR and MAGIg were applied to simulated data sets under various settings. To investigate the performance of the methods under different degrees of differential methylation, the λ parameter was varied as $\lambda = \{0.2, 0.3, 0.4, 0.5, 0.6\}$. To examine the robustness of the methods for different sample sizes, three different replicate numbers (3, 4 and 5 samples) were simulated per group. For FLR, there are B=1000 bootstrap samples used in the test of the fuzzy regression coefficient. The false discovery rate (FDR) is controlled at 5% for each analysis. The methods were compared by calculating the TPR and FPR.

3.3.2. Results. For the simulation with three replicates per group, results for the four methods (FLR, OLR, WALR, and MAGIg) are summarized in Table 3.4 for different values of the methylation difference strength parameter λ. Both the TPR and FPR are reported. The FPR is zero for MAGIg and WALR, but is still small and controlled to be below 5% for OLR and FLR. The TPR for varying degrees of methylation difference λ are displayed in graphical form in Figure 3.5 for each of the four methods and replicate numbers. FLR had the highest TPR value of 0.98 for all λ values. Both WALR and OLR

Table 3.4. Results for a sample size of n=3 per group. True positive rate and false positive rate are given for FLR, OLR, WALR, and MAGIg with various levels of strength of methylation change λ.

CG Three Replicate			
λ	Method	TPR	FPR
0.2	FLR	0.98	0.0014
0.2	OLR	0.92	0.0014
0.2	WALR	0.96	0
0.2	MAGIg	0.84	0
0.3	FLR	0.98	0.0014
0.3	OLR	0.92	0.0014
0.3	WALR	0.96	0
0.3	MAGIg	0.85	0
0.4	FLR	0.98	0.0014
0.4	OLR	0.93	0
0.4	WALR	0.96	0
0.4	MAGIg	0.85	0
0.5	FLR	0.98	0.0014
0.5	OLR	0.93	0
0.5	WALR	0.97	0
0.5	MAGIg	0.85	0
0.6	FLR	0.98	0.0014
0.6	OLR	0.93	0
0.6	WALR	0.97	0
0.6	MAGIg	0.86	0

yielded lower TPR values than FLR, but still over 0.90 for all λ values. WALR always yielded higher TPR than OLR. The MAGIg method did not perform as well, with TPR values between 0.84-0.86.

Results for the simulation with four replicates per group are summarized in Table 3.5 and Figure 3.5. These results are similar to the patterns observed in the three replicate simulation. The FPR is small (less than 0.05) for all methods and is zero for WALR and MAGIg. The FLR, WALR, and OLR performed well in terms of TPR across all λ

Table 3.5. Results for a sample size of $n = 4$ for each group. True positive rate and false positive rate are given for FLR, OLR, WALR, and MAGIg with various levels of strength of methylation change λ.

CG Four Replicate			
λ	Method	TPR	FPR
0.2	FLR	0.99	0.0014
0.2	OLR	0.92	0
0.2	WALR	0.96	0
0.2	MAGIg	0.84	0
0.3	FLR	0.99	0.0014
0.3	OLR	0.92	0.0014
0.3	WALR	0.96	0
0.3	MAGIg	0.85	0
0.4	FLR	0.99	0.0014
0.4	OLR	0.93	0
0.4	WALR	0.96	0
0.4	MAGIg	0.85	0
0.5	FLR	0.99	0.0014
0.5	OLR	0.93	0
0.5	WALR	0.97	0
0.5	MAGIg	0.85	0
0.6	FLR	1.00	0.0014
0.6	OLR	0.93	0
0.6	WALR	0.97	0
0.6	MAGIg	0.86	0

values, with FLR always slightly outperforming WALR, which always outperforms OLR. The MAGIg method again had reduced performance, similar to that in the three replicate simulation.

For the simulation with five replicates per group, the results are given in Table 3.6 and Figure 3.5. For the TPR and FPR, the results are similar to the three and four replicate simulations. All methods have FPR below 0.05. For TPR, the FLR always had the highest value followed by WALR and then OLR. MAGIg was the only method with TPR values below 0.9.

Table 3.6. Results for a sample size of n=5 for each group. True positive rate and false positive rate are given for FLR, OLR, WALR, and MAGIg with various levels of strength of methylation change λ.

CG Five Replicate			
λ	Method	TPR	FPR
0.2	FLR	0.99	0.0014
0.2	OLR	0.92	0
0.2	WALR	0.96	0
0.2	MAGIg	0.84	0
0.3	FLR	0.99	0.0014
0.3	OLR	0.92	0.0014
0.3	WALR	0.96	0
0.3	MAGIg	0.85	0
0.4	FLR	0.99	0.0014
0.4	OLR	0.93	0
0.4	WALR	0.96	0
0.4	MAGIg	0.85	0
0.5	FLR	0.99	0.0014
0.5	OLR	0.93	0
0.5	WALR	0.97	0
0.5	MAGIg	0.85	0
0.6	FLR	1.00	0.0014
0.6	OLR	0.93	0
0.6	WALR	0.97	0
0.6	MAGIg	0.86	0

In conclusion, all methods had a low FPR (<0.05) across all λ values and replicate numbers. All methods except the FLR and OLR actually had no false positives across all settings, but these methods still only had a very small proportion of false positives, well below 0.05. The TPR results for the three, four, and five replicate simulations were all very similar, indicating there was not a clear performance difference for sample sizes in the 3-5 replicate range. Additional studies are needed to investigate a larger variety of replicate

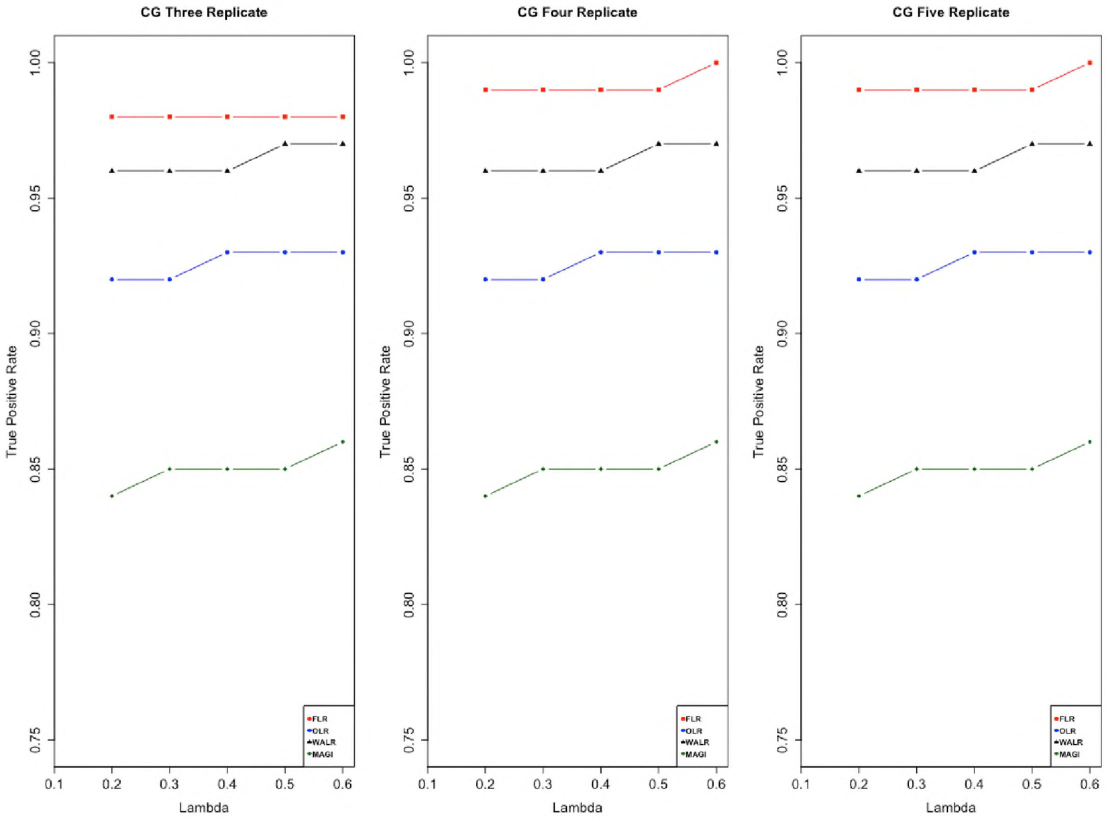

Figure 3.5. True positive rate versus λ level for controlling the degree of differential methylation for each of the four methods (FLR-Red, OLR-Blue, WALR-Black, and MAGIg-Green). Separate graphs are given for 3, 4, and 5 replicates per group.

numbers to better understand the effect of sample size. The simulations in this research indicate a clear performance advantage of FLR and WALR, with FLR having a slightly better TPR across all settings.

3.4. SUMMARY OF SIMULATION STUDIES

This research demonstrates that information from RRBS data sets can be analyzed using the fuzzy logistic regression approach. A FLR method was developed that successfully detects information that cannot be identified by ordinary logistic regression and tests for DMRs between case and control groups. Two other logistic regression approaches (WALR and OLR) were also proposed for DMR testing. All of the proposed methods were compared to the existing method, MAGIg, via simulation studies based on real data. One of the simulation studies utilized plant data to investigate method performance in DMR detection using annotation based regions (genes) under different sequence contexts (CG, CHG, CHH). The other simulation study utilized human data in only the CG context with regions defined based on CG density. Three different replicate numbers (3, 4, 5) were investigated in this study.

An empirical comparison showed that all methods were effective in controlling the false positive rate below the significance level $\alpha = 0.05$ in all simulated data sets. The FLR method consistently had a high true positive rate across all settings compared to the MAGIg, WALR and OLR approaches. The MAGIg method had the lowest TPR across most settings. In the plant data, the OLR approach was better than the WALR for most settings but this was reversed in the human data. Further investigation is needed to understand this difference. One possible explanation could be due to the difference in how the regions are defined since regions in the plant data are based on annotation (genes) while regions in the human data are based on CG density.

For the plant data, the FLR and OLR had consistently high TPR in all three sequence contexts. WALR and MAGIg had the highest TPR (with similar performance to FLR and OLR) for CHH and lowest TPR for CG. This could be due to the difference in correlation between methylation levels of neighboring sites, with CHH having lower correlation levels than CG. For the human data, there was not a notable difference in method performance when comparing data sets with three, four, or five replicates per sample. Larger differences might be observed if additional sample sizes are studied. In all studies, a slight TPR performance improvement is observed for larger values of methylation difference (λ), but the differences are not large.

4. CONCLUSION

The main purpose of this dissertation is to provide a novel statistical framework for identifying differentially methylation regions (DMRs) using reduced representation bisulfite sequencing (RRBS) data. Since predefined regions based on annotation or CG density can offer biologically meaningful interpretation, the proposed methods focus on this type of testing. Section 1 provides an introduction to DNA methylation along with relevant concepts in genetics and epigenetics. The technology used to measure DNA methylation is presented, focusing on RRBS. Previous methods used for DMR testing at predefined regions in RRBS data are reviewed, focusing on two primary methods (M3D and MAGIg). This work builds on the logistic regression approach used in MAGIg.

In Section 2, three additional logistic regression methods are proposed to test for DMRs between conditions of interest at predefined regions. Two of these methods (weighted average logistic regression (WALR) and ordinal logistic regression (OLR)) utilize traditional logistic regression methods that are adapted to analyze DNA methylation RRBS data. The main proposed method is fuzzy logistic regression (FLR), which can be used when the observations are fuzzy numbers. This is the first application of FLR to DNA methylation data. This technique is presented in detail using ideas from Pourahmad (2013) for the FLR model fitting and Lee *et al.* (2015) to test for significance of the logistic regression coefficient.

The FLR method is evaluated along with WALR, OLR, and MAGIg in two simulation studies described in Section 3. Simulation studies in both plants and humans indicate that FLR has consistently high true positive rates that are better than other methods under most settings. WALR and OLR yielded true positive rates between FLR (highest) and MAGIg (lowest) in most settings. Interestingly, all models performed similarly well in the CHH context for the plant data, which has the least correlation between methylation levels of neighboring sites. Further investigation of the effect of correlation between neighboring sites and how to incorporate this into the methods is needed. Additional simulation studies may also help understand how alternate settings such as different replicate numbers, varying noise levels, and CG density may impact the model performance.

Overall, the FLR based on fuzzy set theory seems to be more useful than the logistic regression models based on classic set theory. The FLR method has the potential to be beneficial for cancer research as well as in the pursuit of therapies to combat or prevent lupus, muscular dystrophy and other diseases. For example, since hypermethylation occurs early in colon cancer, the FLR method could be used to detect these DMRs that are an important indicator of potential health problems. Further research in FLR modeling for DMR detection could investigate using a crisp possibility term. Additionally, since the FLR models show promise for methylation data, it is also worth exploring applications of fuzzy methods to other types of genomic data.

REFERENCES

Agresti, A., *An Introduction to Categorical Data Analysis*, John Wiley and Sons, 1996.

Akalin, A., Kormaksson, M., Li, S., Garrett-Bakelman, F. E., Figueroa, M. E., Melnick, A., and Mason, C. E., 'MethylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles,' Genome biology, 2012, **13**(10), p. R87.

Al-Ghamdi, A. S., 'Using logistic regression to estimate the influence of accident factors on accident severity,' Accident Analysis & Prevention, 2002, **34**(6), pp. 729–741.

Baumann, D. D. and Doerge, R., 'MAGI: methylation analysis using genome information,' Epigenetics, 2014, **9**(5), pp. 698–703.

Baumann, D. D., Su, Y., Mendis, I., and Olbricht, G. R., 'Differential methylation methods in multi-context organisms,' in '27th Annual Conference on Applied Statistics in Agriculture,' 2015 .

Benjamini, Y. and Hochberg, Y., 'Controlling the false discovery rate: A practical and powerful approach to multiple testing,' Journal of the royal statistical society: series B (Methodological), 1995, **57**(1), pp. 289–300.

Berglund, E. C., Kiialainen, A., and Syvänen, A.-C., 'Next-generation sequencing technologies and applications for human genetic history and forensics,' Investigative genetics, 2011, **2**(1), p. 23.

Bird, A., 'DNA methylation patterns and epigenetic memory,' Genes & development, 2002, **16**(1), pp. 6–21.

Brettin, T., Davis, J. J., Disz, T., Edwards, R. A., Gerdes, S., Olsen, G. J., Olson, R., Overbeek, R., Parrello, B., Pusch, G. D., *et al.*, 'RASTtk: a modular and extensible implementation of the RAST algorithm for building custom annotation pipelines and annotating batches of genomes,' Scientific reports, 2015, **5**, p. 8365.

Celmiņš, A., 'Least squares model fitting to fuzzy vector data,' Fuzzy sets and systems, 1987, **22**(3), pp. 245–269.

Chang, P.-T. and Lee, E. S., 'A generalized fuzzy weighted least-squares regression,' Fuzzy Sets and Systems, 1996, **82**(3), pp. 289–298.

Dayton, C. M., 'Logistic regression analysis,' Stat, 1992, pp. 474–574.

DeBruyn, J. M., 'Teaching the central dogma of molecular biology using jewelry,' Journal of Microbiology & Biology Education: JMBE, 2012, **13**(1), p. 62.

Diamond, P., 'Least squares fitting of several fuzzy variables,' in 'Preprints of Second IFSA World Congress, Tokyo, Japan,' 1987 pp. 329–331.

Down, T. A., Rakyan, V. K., Turner, D. J., Flicek, P., Li, H., Kulesha, E., Graef, S., Johnson, N., Herrero, J., Tomazou, E. M., *et al.*, 'A Bayesian deconvolution strategy for immunoprecipitation-based DNA methylome analysis,' Nature biotechnology, 2008, **26**(7), p. 779.

Dubois, D. and Prade, H., *Possiblity theory: An approach to computerized processing of uncertainty*, Plenum Press, 1988.

Feinberg, A. P. and Tycko, B., 'The history of cancer epigenetics,' Nature Reviews Cancer, 2004, **4**(2), p. 143.

Feinberg, A. P. and Vogelstein, B., 'Hypomethylation distinguishes genes of some human cancers from their normal counterparts,' Nature, 1983, **301**(5895), p. 89.

Finnegan, E., 'DNA methylation: a dynamic regulator of genome organization and gene expression in plants,' in 'Plant Developmental Biology-Biotechnological Perspectives,' pp. 295–323, Springer, 2010.

Frommer, M., McDonald, L. E., Millar, D. S., Collis, C. M., Watt, F., Grigg, G. W., Molloy, P. L., and Paul, C. L., 'A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands.' Proceedings of the National Academy of Sciences, 1992, **89**(5), pp. 1827–1831.

Gehring, M. and Henikoff, S., 'DNA methylation dynamics in plant genomes,' Biochimica et Biophysica Acta (BBA)-Gene Structure and Expression, 2007, **1769**(5-6), pp. 276–286.

Gu, H., Bock, C., Mikkelsen, T. S., Jäger, N., Smith, Z. D., Tomazou, E., Gnirke, A., Lander, E. S., and Meissner, A., 'Genome-scale DNA methylation mapping of clinical samples at single-nucleotide resolution,' Nature methods, 2010, **7**(2), p. 133.

Gu, H., Smith, Z. D., Bock, C., Boyle, P., Gnirke, A., and Meissner, A., 'Preparation of reduced representation bisulfite sequencing libraries for genome-scale DNA methylation profiling,' Nature protocols, 2011, **6**(4), p. 468.

Hosmer, D. W., Hosmer, T., Le Cessie, S., and Lemeshow, S., 'A comparison of goodness-of-fit tests for the logistic regression model,' Statistics in medicine, 1997, **16**(9), pp. 965–980.

Issa, J.-P. J. and Kantarjian, H. M., 'Targeting DNA methylation,' Clinical Cancer Research, 2009, **15**(12), pp. 3938–3946.

Jaenisch, R. and Bird, A., 'Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals,' Nature genetics, 2003, **33**(3s), p. 245.

Jones, P. A. and Baylin, S. B., 'The epigenomics of cancer,' Cell, 2007, **128**(4), pp. 683–692.

Kao, C. and Chyu, C.-L., 'A fuzzy linear regression model with better explanatory power,' Fuzzy Sets and Systems, 2002, **126**(3), pp. 401–409.

Kim, J., Samaranayake, M., and Pradhan, S., 'Epigenetic mechanisms in mammals,' Cellular and molecular life sciences, 2009, **66**(4), p. 596.

Kleinbaum, K., Kupper, L., Nizam, A., and Muller, K., *Applied Regression Analysis and Other Multivariable Methods, 4e*, Duxbury Press, 2008.

Krueger, F., Kreck, B., Franke, A., and Andrews, S. R., 'DNA methylome analysis using short bisulfite sequencing data,' Nature methods, 2012, **9**(2), p. 145.

Law, J. A., Du, J., Hale, C. J., Feng, S., Krajewski, K., Palanca, A. M. S., Strahl, B. D., Patel, D. J., and Jacobsen, S. E., 'Polymerase IV occupancy at RNA-directed DNA methylation sites requires SHH1,' Nature, 2013, **498**(7454), pp. 385–389.

Law, J. A. and Jacobsen, S. E., 'Establishing, maintaining and modifying DNA methylation patterns in plants and animals,' Nature Reviews Genetics, 2010, **11**(3), p. 204.

Lee, W.-J., Jung, H. Y., Yoon, J. H., and Choi, S. H., 'The statistical inferences of fuzzy regression based on bootstrap techniques,' Soft Computing, 2015, **19**(4), pp. 883–890.

Li, E. and Zhang, Y., 'DNA methylation in mammals,' Cold Spring Harbor perspectives in biology, 2014, **6**(5), p. a019133.

Lim, D. and Maher, E., 'DNA methylation : a form of epigenetic control of gene expression,' The Obstetrician & Gynaecologist, 2010, **12(1)**, pp. 37–42.

Lister, R., O'Malley, R. C., Tonti-Filippini, J., Gregory, B. D., Berry, C. C., Millar, A. H., and Ecker, J. R., 'Highly integrated single-base resolution maps of the epigenome in *Arabidopsis*,' Cell, 2008, **133**(3), pp. 523–536.

Lodish, H., Berk, A., Zipursky, S., Matsudaira, P., Baltimore, D., and Darnell, J., *Molecular Cell Biology 4th edition*, W.H. Freeman, 2000.

Lun, A. T. and Smyth, G. K., 'De novo detection of differentially bound regions for chip-seq data using peaks and windows: controlling error rates correctly,' Nucleic acids research, 2014, **42**(11), pp. e95–e95.

Margot, J. B., Ehrenhofer-Murray, A. E., and Leonhardt, H., 'Interactions within the mammalian DNA methyltransferase family,' BMC molecular biology, 2003, **4**(1), p. 7.

Masser, D. R., Stanford, D. R., and Freeman, W. M., 'Targeted DNA methylation analysis by next-generation sequencing,' JoVE (Journal of Visualized Experiments), 2015, (96), p. e52488.

Mayo, T. R., Schweikert, G., and Sanguinetti, G., 'M3D: a kernel-based test for spatially correlated changes in methylation profiles,' Bioinformatics, 2015, **31**(6), pp. 809–816.

McCullagh, P., ‘Regression models for ordinal data,’ Journal of the Royal Statistical Society: Series B (Methodological), 1980, **42**(2), pp. 109–127.

Meissner, A., Gnirke, A., Bell, G. W., Ramsahoye, B., Lander, E. S., and Jaenisch, R., ‘Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis,’ Nucleic acids research, 2005, **33**(18), pp. 5868–5877.

Meissner, A., Mikkelsen, T. S., Gu, H., Wernig, M., Hanna, J., Sivachenko, A., Zhang, X., Bernstein, B. E., Nusbaum, C., Jaffe, D. B., *et al.*, ‘Genome-scale DNA methylation maps of pluripotent and differentiated cells,’ Nature, 2008, **454**(7205), p. 766.

Olbricht, G., ‘Poster: Statistical Methods for Next-generation Sequencing DNA Methylation Data,’ in ‘Joint Statistical Meetings,’ 2012 .

Pourahmad, S., *Fuzzy logistic regression models with their application in medicine*, LAP LAMBERT Academic Publishing, 2013.

Pourahmad, S., Ayatollahi, S. M. T., Taheri, S. M., and Agahi, Z. H., ‘Fuzzy logistic regression based on the least squares approach with application in clinical studies,’ Computers & Mathematics with Applications, 2011, **62**(9), pp. 3353–3365.

Qian, W., Miki, D., Zhang, H., Liu, Y., Zhang, X., Tang, K., Kan, Y., La, H., Li, X., Li, S., *et al.*, ‘A histone acetyltransferase regulates active DNA demethylation in *Arabidopsis*,’ Science, 2012, **336**(6087), pp. 1445–1448.

Qiu, J., ‘Unfinished symphony,’ Nature Publishing Group, 2006, **441**, pp. 143–145.

Raine, A., Manlig, E., Wahlberg, P., Syvänen, A.-C., and Nordlund, J., ‘Splinted Ligation Adapter Tagging (SPLAT), a novel library preparation method for whole genome bisulphite sequencing,’ Nucleic acids research, 2016, **45**(6), pp. e36–e36.

Robertson, K. D., ‘DNA methylation and human disease,’ Nature Reviews Genetics, 2005, **6**(8), p. 597.

Russo, V. E., Martienssen, R. A., and Riggs, A. D., *Epigenetic mechanisms of gene regulation*, Cold Spring Harbor Laboratory Press, 1996.

Schultz, M. D., Schmitz, R. J., and Ecker, J. R., ‘Leveling the playing field for analyses of single-base resolution DNA methylomes,’ Trends in Genetics, 2012, **28**(12), pp. 583–585.

Serre, D., Lee, B. H., and Ting, A. H., ‘MBD-isolated Genome Sequencing provides a high-throughput and comprehensive survey of DNA methylation in the human genome,’ Nucleic acids research, 2009, **38**(2), pp. 391–399.

Shames, D. S., Minna, J. D., and Gazdar, A. F., ‘DNA methylation in health, disease, and cancer,’ Current molecular medicine, 2007, **7**(1), pp. 85–102.

Slotkin, R. K. and Martienssen, R., ‘Transposable elements and the epigenetic regulation of the genome,’ Nature reviews genetics, 2007, **8**(4), p. 272.

Steyerberg, E. W., Harrell Jr, F. E., Borsboom, G. J., Eijkemans, M., Vergouwe, Y., and Habbema, J. D. F., 'Internal validation of predictive models: efficiency of some procedures for logistic regression analysis,' Journal of clinical epidemiology, 2001, **54**(8), pp. 774–781.

Suzuki, M. M. and Bird, A., 'DNA methylation landscapes: provocative insights from epigenomics,' Nature Reviews Genetics, 2008, **9**(6), p. 465.

Tanaka, H., Uejima, S., and Asai, K., 'Linear regression analysis with fuzzy model,' IEEE Trans. Systems Man Cybern, 1982, **12**, pp. 903–907.

Topal, M. D. and Fresco, J. R., 'Complementary base pairing and the origin of substitution mutations,' Nature, 1976, **263**(5575), p. 285.

Vaillant, I. and Paszkowski, J., 'Role of histone and DNA methylation in gene regulation,' Current opinion in plant biology, 2007, **10**(5), pp. 528–533.

van Eijk, K. R., de Jong, S., Boks, M. P., Langeveld, T., Colas, F., Veldink, J. H., de Kovel, C. G., Janson, E., Strengman, E., Langfelder, P., *et al.*, 'Genetic analysis of DNA methylation and gene expression levels in whole blood of healthy human subjects,' BMC genomics, 2012, **13**(1), p. 636.

Varley, K. E., Gertz, J., Bowling, K. M., Parker, S. L., Reddy, T. E., Pauli-Behn, F., Cross, M. K., Williams, B. A., Stamatoyannopoulos, J. A., Crawford, G. E., *et al.*, 'Dynamic DNA methylation across diverse human cell lines and tissues,' Genome research, 2013, **23**(3), pp. 555–567.

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., *et al.*, 'The sequence of the human genome,' Science, 2001, **291**(5507), pp. 1304–1351.

Voelkerding, K. V., Dames, S., and Durtschi, J. D., 'Next generation sequencing for clinical diagnostics-principles and application to targeted resequencing for hypertrophic cardiomyopathy: a paper from the 2009 William Beaumont Hospital Symposium on Molecular Pathology,' The Journal of molecular diagnostics, 2010, **12**(5), pp. 539–551.

Weaver, I. C., Cervoni, N., Champagne, F. A., D'Alessio, A. C., Sharma, S., Seckl, J. R., Dymov, S., Szyf, M., and Meaney, M. J., 'Epigenetic programming by maternal behavior,' Nature neuroscience, 2004, **7**(8), p. 847.

Wu, H.-C., 'Linear regression analysis for fuzzy input and output data using the extension principle,' Computers & Mathematics with Applications, 2003, **45**(12), pp. 1849–1859.

Yen, K. K., Ghoshray, S., and Roig, G., 'A linear regression model using triangular fuzzy number coefficients,' Fuzzy sets and systems, 1999, **106**(2), pp. 167–177.

Zadeh, L. A., 'Fuzzy sets,' Information and control, 1965, **8**(3), pp. 338–353.

Zhong, X., Hale, C. J., Law, J. A., Johnson, L. M., Feng, S., Tu, A., and Jacobsen, S. E., 'DDR complex facilitates global association of RNA polymerase V to promoters and evolutionarily young transposons,' Nature structural & molecular biology, 2012, **19**(9), p. 870.

Zilberman, D., Gehring, M., Tran, R. K., Ballinger, T., and Henikoff, S., 'Genome-wide analysis of *Arabidopsis thaliana* DNA methylation uncovers an interdependence between methylation and transcription,' Nature genetics, 2007, **39**(1), p. 61.

Zimmerman, H.-J., *Fuzzy set theory and its applications*, Kluwer, 1996.

VITA

In 1998, Tarek M. Bubaker Bennaser received a B.S. in Mathematics from Garyounis University, Libya. In 2008, Tarek received his M.S. in Pure Mathematics from Zawiya University, Libya. In May 2009, he joined Garyounis University, Libya, as a lecturer in the Mathematics and Statistics Department. In 2014, Tarek received his M.S. in Applied Mathematics from Missouri University of Science and Technology, and in December 2020 he received his Ph.D. in Mathematics with Statistics emphasis from Missouri University of Science and Technology.